



---

## 2-FUN

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

### **FP6 Project-2005-Global-4**

**Integrated Project - Contract n°: 036976**

---

## **– D1.13 COMPARISON OF METHODS OF MATHEMATICAL GNOSTICS WITH STATISTICAL ONES IN APPLICATIONS –**

Due date of delivery: *31/01/2011*

Actual submission date: *31/01/2011*

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) |        |
| <b>Dissemination Level</b>  |        |
| <b>PU</b>   | Public |



## Document Information

**Document Name** D1.13 Comparison of methods of mathematical gnostics with statistical ones in applications  
**ID** D1.13\_Comparison\_of\_uncerainty\_methods.doc  
**Revision** Version 3  
**Revision Date** 11/03/2011  
**Author** P. Kovanic

## Approvals

|                    | Name         | Company | Date       | Visa         |
|--------------------|--------------|---------|------------|--------------|
| <b>Author</b>      | P. KOVANIC   | IPH     | 31/01/2011 | P. Kovanic   |
| <b>WP Leader</b>   | A. MARCOMINI | UNIVE   | 31/01/2011 | A. Marcomini |
| <b>Coordinator</b> | F. BOIS      | INERIS  | 31/01/2011 | F. Bois      |

## Documents history

| Revision  | Date       | Modification                | Author       |
|-----------|------------|-----------------------------|--------------|
| Version 1 | 20/01/2011 | First version               | P. Kovanic   |
| Version 2 | 31/01/2011 | Layout and format revisions | E. Giubilato |
| Version 3 | 31/01/2011 | Final version               | F. Bois      |



## Contents

|   |           |
|---|-----------|
| <b>Introduction and references</b>  | <b>4</b>  |
| <b>1. Part I. – Comparison of theoretic backgrounds of statistics and mathematical gnostics</b> | <b>4</b>  |
| 1.1 On puzzles of quantitative uncertainty  | 4         |
| 1.2 Galilean/Newtonian roots of statistics  | 6         |
| 1.3 Riemannian and relativistic roots of mathematical gnostics                                  | 7         |
| <b>2 Part II.: Comparison of results in applications</b>  | <b>8</b>  |
| 2.1 A review of previous comparisons  | 8         |
| 2.1.1 Remarkable applications in substitution for statistics                                    | 8         |
| 2.1.1.1 Fatigue cracks of locomotive's driving axes   | 8         |
| 2.1.1.2 Unstable quality of a fertilizer  | 9         |
| 2.1.1.3 A technological problem with suspension springs of heavy trucks TATRA                   | 9         |
| 2.1.1.4 The quality of the Caprolactam  | 9         |
| 2.1.1.5 The Cleanroom Problem   | 11        |
| 2.1.1.6 Other successful applications   | 12        |
| 2.1.2 Already described direct comparisons of methods   | 13        |
| 2.2 Recent direct comparisons of methods  | 14        |
| 2.2.1 Robust trends of contamination  | 14        |
| 2.2.2 Brain tumors in North Moravia   | 15        |
| 2.2.3 Identification of the stackloss model   | 17        |
| 2.2.4 Historical data on fertility in Switzerland   | 18        |
| 2.2.5 Stock market predictions  | 19        |
| <b>3 The normality problem of uncertain events</b>  | <b>20</b> |
| 3.1 What is to be really normal?  | 20        |
| 3.2 Statistical setting the bounds of the normal/reference range                                | 20        |
| 3.3 Empirical reference range in clinical practice  | 21        |
| 3.4 Gnostic bounds of normality of uncertain events   | 21        |
| 3.4.1 Example 1: Normality of testosterone level in postmenopausal women                        | 22        |
| 3.4.2 Example 2: Normality of estradiol level in postmenopausal women                           | 24        |
| <b>4 Conclusions</b>  | <b>25</b> |



## Introduction and references

Mathematical gnostics is a methodology of data treatment based on the gnostic theory of individual uncertain data and small samples. A complete review of this theory along with the program package usable in the free open source environment of the R-project is available on [www.math-gnostics](http://www.math-gnostics). This methodology is a non-statistical alternative to both classical and robust statistics suitable for application to strongly uncertain data. A brief comparison of theoretical backgrounds of statistical and gnostic approaches along with comparisons of the results of applications are presented to demonstrate not only legitimacy, but also efficiency of the alternative approach.

### 1. Part I. – comparison of theoretic backgrounds of statistics and mathematical gnostics

#### 1.1. On puzzles of quantitative uncertainty

Quantitative recognition of the world is a mapping of structures of real quantities into structures of numbers (**quantification**), i.e. mapping of Nature into mathematics. To serve to people's decision making, the inverse mapping must be performed, the **estimation**. Quantification results in **data**, obtained by measuring and/or counting. Problem is that real processes and their quantification can be uncertain. Decision making based on uncertain data can thus be risky. The fight against uncertainty spans for centuries, at least from the time of **Luca Pacioli** (1445-1517). It includes not only development of statistics, but a much broader front recently represented by the "movement" IPMU (International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems), organizing large-scale meetings every two years (<http://www.mathematik.uni-marburg.de/~ipmu2010/>.) There were 23 different approaches to uncertainty discussed at the occasion of IPMU2010 in Dortmund. This profusion of models of uncertainty documents the difficulty of the topic and the non-existence of a generally accepted and unique paradigm of uncertainty. The state of art can be rather formulated as "running competition of paradigms of uncertainty".

This is fully relevant for the field of HRA (Health Risk Assessment), where the risks are priced on the life/death level. Sharpness of this application calls for fundamental problem setting starting with the most "elementary" questions:

- 1) A result of measuring the actual quantity  $q$  was  $Q$  because of uncertainty. What was the error of this data item? The answer of the classical statistics " $Q$  minus  $q$ " invokes the "childlike" reaction: WHY? The difference is application of the Lebesgue's measure of distance applicable to a subspace of the Euclidean space. This means, that validity of Euclidean geometry has been assumed. But this is geometry of a non-curved space. Generally accepted opinions exist (accepted e.g. by methods of robust statistics), that spaces of uncertain data should be considered as points of some curved spaces. However, accepting this opinion, an even more complex problem is arising: what kind of geometry is the "proper" or "natural" one of uncertainty?



- 2) What should be the composition law for two errors  $E_1$  and  $E_2$ ? The “standard” answer “ $E_1 + E_2$ ” again raises the objection WHY? This is another application of Euclidean geometry. But this is the “ordinary” way of getting the “mean” value.
- 3) To evaluate the spread of data (by the variance), squares of data errors are composed in statistics additively as well as products of errors composed to evaluate the covariance. This is equivalent to calculation of the distances and angles in a Euclidean multidimensional space. What is the reason for considering the uncertain data as points of the Euclidean space? One of Euclidean axioms says that a straight line connecting two points exists to an arbitrary pair of points. Moreover, this line segment should be the shortest path between the points. But to think on such a design in a real space is in a conflict with the finite speed of light and with the theory of gravitation.
- 4) To evaluate risk, one needs probability. In his book<sup>1</sup> T.L.Fine analysed seven theories of probability existed at his time to come to conclusion:  
*“ . . . 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?”*  
The HRA needs probability. What is to be used?
- 5) Uncertainty is traditionally interpreted as a counterpart of information. If so, then amounts of uncertainty could be measured by the corresponding amounts of information. But this could be done only when formulae of amount of information and of its interdependence with uncertainty would be available. However, the Shannon’s idea of reducing this problem to application of the negative Boltzmann’s entropy is not suitable in HRA, because its application would need a complete probabilistic description of the treated data, which is generally not available.
- 6) There is a fundamental limitation of statistics dwelled in limited applicability of the Central Limit Theorem: it assumes that the considered data have a mean and a second statistical moment. It can be shown, that this limitation prevents to statistically treat some strongly dispersed data.
- 7) Number of data available for HRA is limited because of the difficulties of measurements and high costs of both field and laboratory work necessary to get the data for the mathematical treatment. The economics of information obtained from data is therefore an unavoidable problem in HRA solvable only when resulting information is maximized. Optimization criterion of the data treatment is thus obvious: maximization of information, minimization of uncertainty. Measurement of these quantities is a must, but it requires a consistent theoretical fundament.

---

<sup>1</sup> Fine T.L., Theories of probability; an Examination of Foundations, Academic Press, New York and London (1973).



According to B. Blažek<sup>2</sup>, science is developing by explicitation of concealed assumptions. This statement is not related to possible dishonest actions of creators of scientific theories, but to historical development of sciences. Euclid surely could not conceal the fact of limited speed of light, because at his time nobody had an idea of something like this. The Euclidean axioms were subjected to critical revision at nineteenth century leading to creation of alternative and more general geometries. But it was physics, which was able to reveal the hidden Euclidean assumption on possibility of synchronous events in a large space and to introduce the Lorentz-invariant space-time geometry.

## 1.2 Galilean/Newtonian roots of statistics

Statistics is a mathematical science. Attempts to justify its basic assumptions mathematically are therefore legal. It has been shown<sup>3</sup>, that an isomorphic linear mapping exists between the statistical dyad “data error, squared error” and the dyad of classical mechanics “momentum, kinetic energy”. Such a relation can seem only formal, but there are fundamental consequences of it to be recognized:

- 1) Such a mapping Mechanics  $\leftrightarrow$  Statistics appears to be invariant under the Euclidean transformations (shifts and orthogonal rotations).
- 2) Classical mechanics is based on the much older Euclidean geometry simply because it “worked” on practice of physics. The considered mapping thus justifies application of Euclidean measuring of the statistical errors.
- 3) Total momentum and energy of a system of free particles is aggregated additively by moments and energies of individual particles. This motivates additive composition of statistical errors and of their squares to preserve the said linear mapping of systems.
- 4) The additive aggregation law of moments and energies is supported by the Energy-momentum Conservation Law of classical mechanics, the validity of which for sufficiently slow movements was never falsified.

The requirements of the frequently used estimating method BLUE (the Best Linear Unbiased Estimate) can be thus “translated” as the conditions of “zero total moment of errors” and “minimum total energy of errors”. It is highly plausible, that this was the way of thinking of the classical “inventors” of the Least Squares Method, who were not only mathematicians, but also physicists.

Accepting this point of view one admits, that classical mechanics with its Euclidean roots justifies not only Euclidean measuring and additivity of statistical errors, but also the additive composition of squared errors. On the other hand, Euclidean geometry is applicable only to spaces with no curvature and classical mechanics is valid only for low velocities. These facts justify the point of view, that the applicability of classical statistics is seriously limited not only practically, but also theoretically.

---

<sup>2</sup> Bohuslav Blažek (1942-2004): Czech philosopher and writer.

<sup>3</sup> Kovanic P.: Gnostic theory of uncertain data, Doctor (DrSc.) thesis (in Czech), Institute of Information Theory and Automation of the Czechoslovak Academy of Sciences (1990), 152 pp., 9 figs.



### 1.3 Riemannian and relativistic roots of mathematical gnostics

Fundamental features of mathematical Gnostics are derived already in its first part, theory of individual uncertain data. Unlike statistics representing the contamination of the true value of a data item by uncertainty by a one-dimensional additive model, the gnostical model is bi-dimensional. The first axiom assumes that both true and uncertain components of the data item satisfy the conditions laid down on elements of a commutative Abelian group subjected to the same structure operation inside and between the groups. This model is a two-dimensional generalization of that of the classical von Helmholtz's measurement theory. Reformulation of the axiom in ordinary language: "considered data are obtained by proper counting or measuring procedures". Consequences of this axiom are far reaching:

- 1) An individual data item contaminated by an uncertainty can be modeled by an element of the bi-algebra endowed by Minkowskian metric. This element is called the **pair number**.
- 2) The virtual path from the true to observed value can be modeled as a branch of Minkowskian circle, the diameter of which is the invariant of virtual movement.
- 3) This ("quantification") path is the extreme line of the space: its length is the **maximum** among the alternative lines connecting the same end points.
- 4) When looking for the best (**minimum** length), "estimating" back path from the observed to true data value, one finds the Euclidean circle, the points of which are represented by complex numbers.
- 5) The double ("quantifying" and "estimating") representation of an uncertain data item enables two pairs of nonlinear characteristics of data uncertainty to be defined: quantifying and estimating **irrelevance** and **data weight**.
- 6) The irrelevances can be shown to represent data errors measured by using certain Riemannian geometries of curved spaces.
- 7) The data weights can be interpreted as equivalents of the change of the Clausius' classical **entropy** caused by the data uncertainty measured by using Riemannian geometries.
- 8) Looking for sources of entropy fields over the Minkowskian and Gaussian spaces and integrating these sources along the (quantifying and estimating) virtual paths, formulae of **information** change caused by uncertainties in a data item are derived along with the equation quantifying the conversion of information into entropy within the quantification and backwards by the estimation.
- 9) Application of measuring the uncertainty by irrelevance and data weights results in two kinds of natural **robustness**: to outliers (estimating case) and to inliers (quantifying measures).
- 10) Quantifying and estimating virtual paths form the **Ideal Gnostic Cycle** of the transformation of the uncertain data. Idealness of this cycle consists of maximization of the entropy increase during quantification and maximization of information yielded by estimation when following the virtual paths by using the Riemannian formulae of mathematical gnostics. Extreme features of this cycle are proved.
- 11) Existence of the one-to-one linear mapping between the quantifying pair (irrelevance, data weight) and the pair (momentum, energy) of a free relativistic particle is proved. This mapping is invariant under Lorentz's transformations.



- 12) This mapping can be interpreted as the support for the additive composition of irrelevances and data weights, i.e. for the second gnostic axiom determining the way of aggregating the gnostic measures of data uncertainty.
- 13) Two interesting ideas potentially useful for physics can be drawn from the mathematical gnostics:
  - A) The isomorphism between gnostic model of an uncertain event and a moving relativistic particle proves, that information/uncertainty of measurements plays a role of the fifth dimension of a relativistic space-time model. Identification of a moving object is parameterized by its relative velocity with respect to the observer. However, errors in measuring object's velocity are reflected by observed parameters like actual changes of the velocity.
  - B) The probability of an individual data item is a real linear function of the data irrelevance, which quantifies the data error/uncertainty during the estimating phase of the Ideal Gnostic Cycle. However, there also exists its "double" quantifying irrelevance represented by a "pair" number, which corresponds to a point of Minkowskian space. A linear function of this irrelevance can be interpreted as improbability. The mentioned isomorphism allows thus attachment of the improbability values to a moving particle to quantify its uncertainty. The measuring of improbability is Lorentz-invariant. This might be interesting for quantum mechanics.
- 14) It can be shown easily, that in cases of small relative errors of data, the gnostic characteristics of data uncertainty converge to statistical ones. The robust quantities irrelevance and data weight lead in limit of very small errors to the same results like the statistical error and variance. The gnostic curved space of uncertain data is approximated in statistics by a tangential plane with the Euclidean metric. Gnostic model can be thus viewed as a generalization of the statistical model from this point of view.

## **2. Part II.: Comparison of results in applications**

### **2.1. A review of previous comparisons**

The novelty and unusual mathematical nature of the new theory were leading to requirements of tests by applications all the time from its origin to demonstrate its capacity. There were two ways of getting a support for the developing approach:

- a) To help in solving problems of industrial producers and other users of statistics disappointed by failing statistical methods to treat their data.
- b) To directly compare results of both approaches by application to the same data.

#### **2.1.1. Remarkable applications in substitution for statistics**

##### **2.1.1.1. Fatigue cracks of locomotive's driving axes**

Czechoslovak heavy industry was exporting many products into USSR in the past. The Prague machinery factory ČKD was delivering heavy locomotives for the hard conditions of the trans-Siberian magistral. These machines were favored for their reliability till late eighties, when cracks of driving axes started to appear. The research institute of the ČKD was



charged with the task of finding causes of the failures and of proposing a remedy. Locomotives were periodically lifted and their drive gear tested. These measurements were expensive and their (rare) results chaotic in such a degree that statistical methods were failing even when applied by statistical experts.

A careful application of gnostic distribution functions enabled not only identify but also predict the process of fatigue cracking and reliably decide on the remedy actions. Comparison of time series of the distributions led to determination of the time, when problems started. It coincided with introduction of a simplification in production technology of the gear. The problem was solved.

#### **2.1.1.2. Unstable quality of a fertilizer**

Quality of a kind of fertilizer produced by the Spolana chemical factory was dependent on manipulations in postproduction phase. Technology prescribed a certain time necessary for stabilization of chemical processes continuing to run after finishing the production. Nonobservance of rules led to high corrosion of the tools and machines of farmers. Frequent claims on lowering quality forced the department of quality assessment to start investigation of the causes. After failures of statistical methods, gnostic robust estimates revealed the cause by comparison the time series of quality. Significant lowering of quality was detected with products, which were delivering from the store after the night shift. A revision has shown that it resulted from nonperformance of the master of the night shift, who was not respecting the prescribed time order of warehousing of the fertilizer.

#### **2.1.1.3 A technological problem with suspension springs of heavy trucks TATRA**

Moravian producer of air-cooled TATRA heavy trucks became famous by repeated victories of his trucks in Dakar races. However, repeated cracks of suspension springs forced the factory's laboratory to return to some variants of production technology by testing the life time of the spring by many cycles of periodical loading. Such tests are very time intensive, because some springs crack only after millions of cycles, while the other "survive" as long, that stopping the test is necessary. Tests are therefore expensive and data are rare. When statistical methods of treating these small data samples failed, gnostic methods were adapted to this problem and satisfactorily applied. The best variant of technology was found.

#### **2.1.1.4 The quality of the Caprolactam**

Caprolactam is an important chemical product used as a raw material in the textile industry. Some years ago, the quality control department of a producer of Caprolactam initiated a study to identify the factors of production that have the main impact on the quality of this product. The behavioral relationships between the variables traditionally used to measure quality and the main quality indicator, the light absorbance, were not known with sufficient confidence to justify the setting of control thresholds for important variables. Caprolactam was produced in batches and measurements of nine quality indicators from 39 batches were provided as inputs to the study. The variables are listed in Tab.1:



| No. | Symbol | Name                  | Variability |
|-----|--------|-----------------------|-------------|
| 1   | AB     | Light absorbance      | Continuous  |
| 2   | PN     | Permanganate number   | Continuous  |
| 3   | VA     | Volatile alkalis      | Continuous  |
| 4   | CH     | Color by Hazen        | Continuous  |
| 5   | AL     | Alkaescence           | Continuous  |
| 6   | MC     | Moisture content      | Continuous  |
| 7   | SP     | Solidification point  | Dichotomous |
| 8   | AR     | Annealing rest        | Dichotomous |
| 9   | MI     | Mechanical impurities | Dichotomous |

Tab.1: Variables determining the Caprolactam's light absorbance

Only variables 1 through 5 were continuous; the rest were dichotomous: only two levels could be reliably distinguished: 'high' or 'low.' Limitations as to permissible values for absorbance, *AB* (0.6), were set by the end user, and the permanganate number, *PN*, played the next most important role in the manufacturing process.

The analysis was intended to explore the impact of these individual indicators on the absorbance. Application of statistical methods appeared to be unsuitable in solving the problem. The advantage of gnostic distribution function (their ability to estimate the conditioned probability distributions) was used to come to the surprising conclusions, that only the dichotomous variable *AR* was the decisive factor: its value 'low' could really warrant the required quality. It can be seen in Fig.1 (cited from the mentioned book on gnostic theory, where it had No.26.14 and where more details and results on this task can be found), that the guaranty of low *AR* was realized with a large reserve.

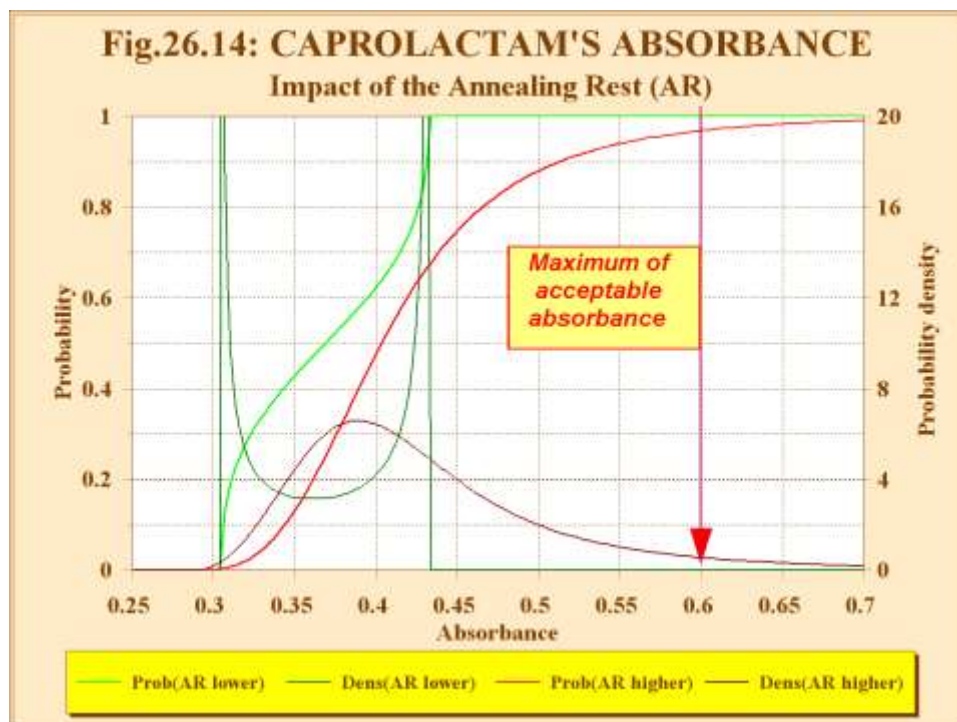


Fig.1: Two conditioned probability and density functions of the Caprolactam batches

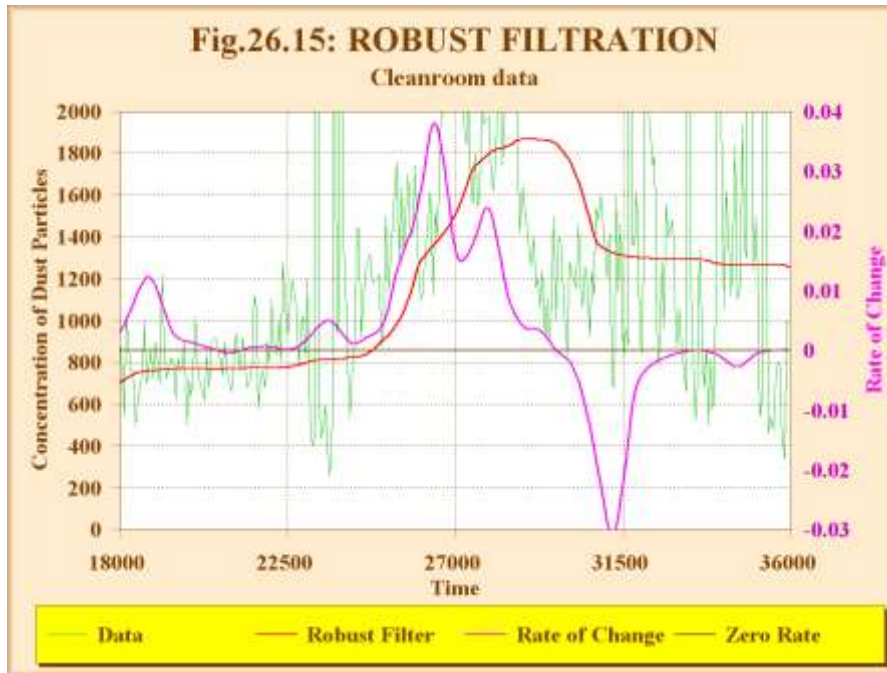


It might be interesting to note, that only 9 batches from all 39 existed with the 'low' value of the variable  $AR$ . The green probability and density distributions were estimated by using these 9 values. Gnostic global distribution functions have been applied. Two important quantities were estimated as a "by-product": The lower and upper bounds of domains (ordinarily denoted by  $LB$  and  $UB$ ) of all the probability and density functions. The lower bound is the quantile, at which both probability and density functions reach the zero probability value. At the upper bound, the density is zero and probability reaches the value of 1. It means that probability of a value crossing the interval  $(LB, UB)$  is zero.

### 2.1.1.5 The Cleanroom Problem

For some manufacturing processes, the quality of the product depends on many factors, among which an important role is played by the environmental conditions that prevail, where the work is performed. Many procedures require extreme cleanliness of the air in the workplace. This is necessary not only in operating rooms, food processing plants or pharmaceutical production, but also in the high-tech industry. Experience has shown that the reliability of microchips depends on the concentration of airborne micro-contamination particles in the rooms, where uncovered chips are manipulated. Such a room is called a *cleanroom*. These special needs are incorporated in the building's plans. There also are requirements to a precise measuring and information system. These problems were met by the American firm DEC (Digital Equipment Corporation, later taken over by another company) when working on the project of a new factory in Massachusetts. To monitor the cleanness of the air, a number of Wilson chambers and laser equipments were projected. However, problems appeared with data filtration and interpretation because the very complex behavior of the airborne micro-contamination particles: the more or less stationary 'noise' generated by many small particles was disturbed by sudden very strong excursions probably caused by a clump of particles clustered together. These bursts were occurring and then randomly ceasing without affecting the background level. The difficult issue to be confronted was to decide, whether these were a very short term temporary disturbances or the start of a dangerous rise in the particles' density. The DEC's bureau in Vienna attempted to solve the problem by an approach based on fractal methodology. In parallel with this effort, a research project was sponsored by the DEC based on gnostic methodology. A decision was made to solve the main tasks by using fast, but robust gnostic recursive filters running in-line. An example of its activity is reproduced in Fig.2 from Fig.26.15 of the mentioned book.

The thin green line shows the unfiltered output of a particle counter, while the thick red line is the output of the gnostic filter. Note, that the filter smoothes even large temporal excursions, but it is sensitive to and reacts in a timely manner to changes in the actual level of the process. This provides reliable information sufficient to satisfy tasks of a warning system. The magenta line in Fig.2 represents the rate of change of the robust filter (its first derivative) obtained by application of the differentiating linear operator (see sections 18.4.1 and 18.4.4 of the book). This output is both robust and sensitive enough to support the automatic decision making on the degree of a danger. Should the filtered detectors' signals fell below a threshold, an auto-diagnostic signal of the monitor would be triggered.



**Fig.2: Robust filtration and differentiation of cleanroom data**

#### 2.1.1.6 Other successful applications

There were more successful applications of gnostic methods outperforming the statistical methods in applications to real data worth of mentioning:

- 1) Gnostic methods were and still are being 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 time series of aerosols.
- 2) Robust models of factors influencing the vitality of living cells cultivated in spinners were tested in a biomedical research project.
- 3) 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.
- 4) Many important economic analyses:
  - a) Analyses of processes of Czechoslovak economic transformations.
  - b) Re-appreciation and prediction of share prices.
  - c) Analysis of the interbank on-line financial exchange market.
  - d) Analysis of dirty financing.
  - e) Mathematical rating of firms.
  - f) Advanced financial statement analysis.
  - g) Marketing study of European car market.Some of these studies are reviewed in the book on mathematical gnostics.
- 5) Analysis of job stresses.
- 6) Analysis of Czech historical coins from the 15<sup>th</sup> century.
- 7) Two military projects.
- 8) Analysis of data from an international inter-laboratory geological survey.
- 9) Treatment of data from monitoring of pollutions in air and waters of Czech Republic.
- 10) Analysis of data from environmental monitoring, bio-monitoring and children morbidity surveys in Poland.



## 2.1.2. Already described direct comparisons of methods

Several direct comparisons of data by application of both statistical and gnostic methods were already presented<sup>4</sup>:

- A) Comparison of robust estimators of location applied to historical experiments of Newcomb and Michelson resulted in following conclusions:
  - Eleven statistical and robust statistical methods gave substantially different results.
  - Gnostic estimates coincided with the mean of the 11 statistical methods.
- B) Comparison of gnostic distribution functions with seven statistical non-parametric methods of estimating the probability density available in R-project: it resulted, that gnostic distributions provide substantially more information on the data samples, on their structure and bounds.
- C) Comparison of robust estimation of correlation coefficients by the gnostic method with 8 methods of robust statistics available in R-project was performed in application to contamination data of six rivers of Czech Republic. Results were analogous to the case of location estimators: different statistical methods gave substantially different results, but their mean coincided with the gnostic results.
- D) Eleven statistical robust regression models of the Iterated Weighted Least Squares type (M-estimators) were compared in application to the problem of impacts of organic pollutants in rivers on toxicity by means of the gnostic method. The gnostic method appeared to warrant the maximum statistical power of probability tests.
- E) Estimation of the left-censored data by six primitive (frequently used) and one theoretically based statistical method was compared with the gnostic method<sup>5</sup>.  
The simple statistical methods were giving erroneous data estimates distorting the sample's distribution functions over the whole data range. The theoretically founded Kaplan-Meier method gave results well coinciding with gnostic results, but only in discrete points, because it does not provide a smooth distribution function. It cannot therefore reliably estimate the lifetime.
- F) Comparison of hypotheses testing was discussed on tests of impacts of chemical industry on contamination of blood of citizens. Unlike gnostic distributions derived from data without any assumptions, the statistical tests are subjective due to necessity to assume the type and form of the probability distributions of tested events. Moreover, they do not allow to estimate the bounds of the distribution domain and are not robust. Reliability of statistical tests is therefore doubtful.

---

<sup>4</sup> Extended report on review and tests of methods for filling data gaps, Kovanic P. and Ocelka T., Deliverable of the 2-FUN project ID D1.9 Extended\_report.docx (2009), 40pp.

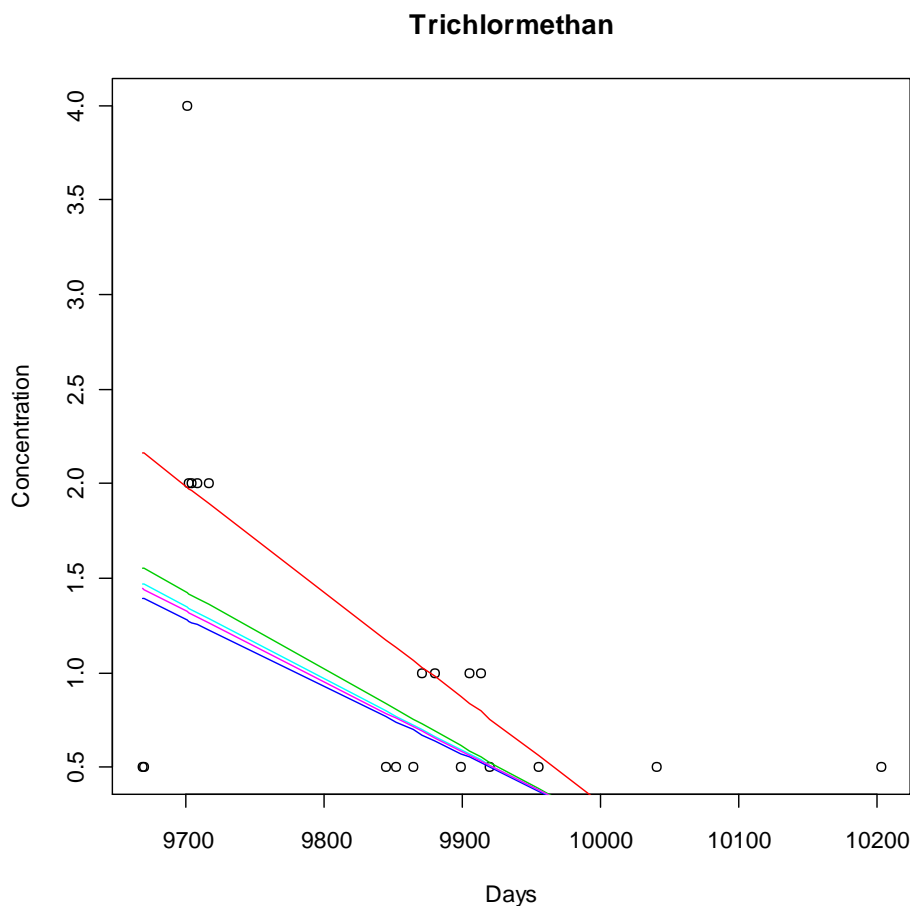
<sup>5</sup> Kovanic P. Ocelka T., Ciffroy P.: An alternative approach to handle non-detectable values in datasets obtained in environmental and health monitoring, to be published.



## 2.2. Recent direct comparisons of methods

### 2.2.1 Robust trends of contamination

As many districts of Stuttgart, the district of Feuerbach is constricted by severe soil and groundwater contamination generated over decades of industrial and commercial use. Due to the structural change many former industrial sites now are converted into service and residential use. Numerous single site investigation and remediation activities took place in Feuerbach having been carrying out since 1984. Thus 300 contaminated sites were identified. 193 of them are polluting or potentially polluting the groundwater by chlorinated hydrocarbons (CHC), which are known to generate long plumes. Problems connected with the contamination are being treated within the framework of the project FOKS ([www.foksproject.eu](http://www.foksproject.eu)). More than 40000 measurements are available recently and the problems with their interpretation are being considered. One of important aspects is the time trend of the contamination. An example can be used for discussion on data treatment methods (Fig.3).



**Fig.3: Comparison of estimated time trends of concentration in a well**

A special feature of these data consists in rounding off the small (and uncertain) data values to one digit. This along with high data variability makes the estimating of the trends difficult.



A review of measurements of trichlormethan in a well in dependence on the number of days since 1.1.1980 is presented in Fig.3. Gnostic function GWLS was applied to these data to compare the results with the robust statistical methods available in R-environment.

| Method          | R-square | MeanW | MErr  | MAErr | MsqErr | Line in Fig.3 |
|-----------------|----------|-------|-------|-------|--------|---------------|
| <b>Gnostic</b>  | 0.97     | 0.60  | -0.07 | 0.25  | 0.49   | Red           |
| <b>OLS</b>      | 0.26     | 1     | 0     | 0.74  | 0.93   | Green         |
| <b>Huber</b>    | 0.42     | 0.98  | 0.03  | 0.72  | 0.89   | Light blue    |
| <b>Hampel</b>   | 0.34     | 0.99  | 0.02  | 0.73  | 0.91   | Dark blue     |
| <b>Bisquare</b> | 0.47     | 0.96  | 0.02  | 0.71  | 0.87   | Magenta       |

**Tab.2: Comparison of linear trends estimated by different methods**

The symbols of statistics in Tab.2:

R-square ... the ratio of the weighted variance explained by the model and of the total weighted variance of the dependent variable.

MeanW ... mean value of weights of individual equations of the regression models resulting from iterations.

MErr ... the mean residual model's error.

MAErr ... the mean absolute residual error of the model.

MsqErr ... the mean square residual error.

The ordinary least squares method (OLS) provides the minimum variance unbiased estimate (MErr=0). The condition of zero mean error prevents lowering the variance below a certain value determined by the Cramer-Rao inequality. Estimates satisfying this condition are called efficient. This name is a little bit confusing with respect to the general sense of the word, because – as known – there are “more efficient” estimates giving a less variance because of not insisting on the unbiasedness. As shown in Tab.2, the gnostic estimate is biased but its mean error is very small in comparison with the MAErr and MsqErr. But the small bias and a low mean weight help in reducing the gnostic weighted MAErr to nearly a third of that of the statistical methods.

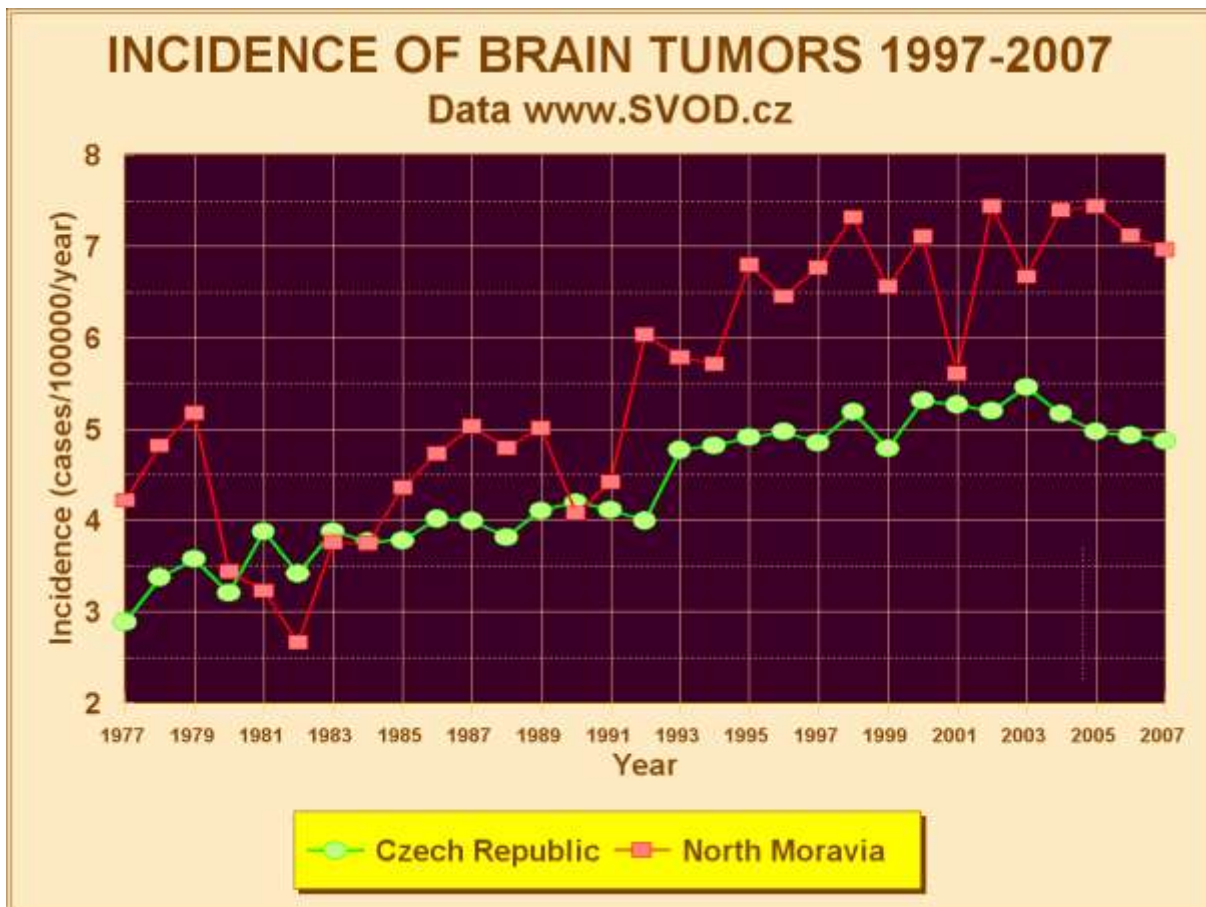
This evaluation was based on the statistical measures of quality, which – unlike the statistical case - are not optimized by the gnostic method. Using the gnostic criteria of quality would speak even more in favor of the gnostic method.

### 2.2.2. Brain tumors in North Moravia

North Moravia is an eastern region of the Czech Republic with a rich industrial tradition (coal mines, metallurgy, machinery production). Environmental data indicate health danger higher than that of the whole country. So, the incidence of brain tumors in this region exceeded that of the whole country, as observed during 1977-2007 (Fig.4). These results call for a thorough quantification. A probabilistic test of the hypothesis „The incidence of brain tumors in North



Moravia exceeds that of the whole Czech Republic“ could be the right way of quantification. Moreover, it can be used for a direct comparison of the gnostic approach with the statistical ones.



**Fig.4: Comparison of incidence of brain tumors**

Consider the statistical test of the hypothesis: Assume logarithmic normality of distributions, estimate samples' mean and standard deviations and apply the Student's t-test of coincidence of the means of the samples. The 95% confidence interval for means is (0.558, 1.710). The actual difference of means is 1.152, hence means differ. The test power would be 0.999, but only in the case of actual normality, which is doubtful at least because of impossibility of infinite incidence of tumors, which could be reached with a non-zero probability according to the log-normal distribution. There also are doubts raised by the actual form of the distribution function.

Gnostic test of the hypothesis performed by the global gnostic distributions estimated by the function *GNDF* substituted into the function *TestHyp* gave results illustrated by Fig.5:

- ❖ The probability distributions of the incidence cannot be viewed as log-normal ones, they have finite domains and their densities have nearly uniform character.
- ❖ Incidence of tumors in North Moravia **exceeds** that of the whole Czech Republic in 52.4% of population by 2.23 cases/10<sup>5</sup>/year.
- ❖ The danger from the interval (2.88, 5.48) cases/10<sup>5</sup>/year threatens to 39.5% of North Moravian population and to 91.9% to the people of the whole country.



- ❖ The danger in the North Moravia is *less* than that of the whole country only in 8.1%.

There also are **almost sure** results of the test, i.e. statements with probability 1:

- ❖ Brain tumors incidence less than 2.66 and exceeding 7.70 cases/10<sup>5</sup>/year should not be expected in North Moravia.
- ❖ Brain tumors incidence less than 2.88 and exceeding 5.48 cases/10<sup>5</sup>/year should not be expected in the Czech Republic as a whole.

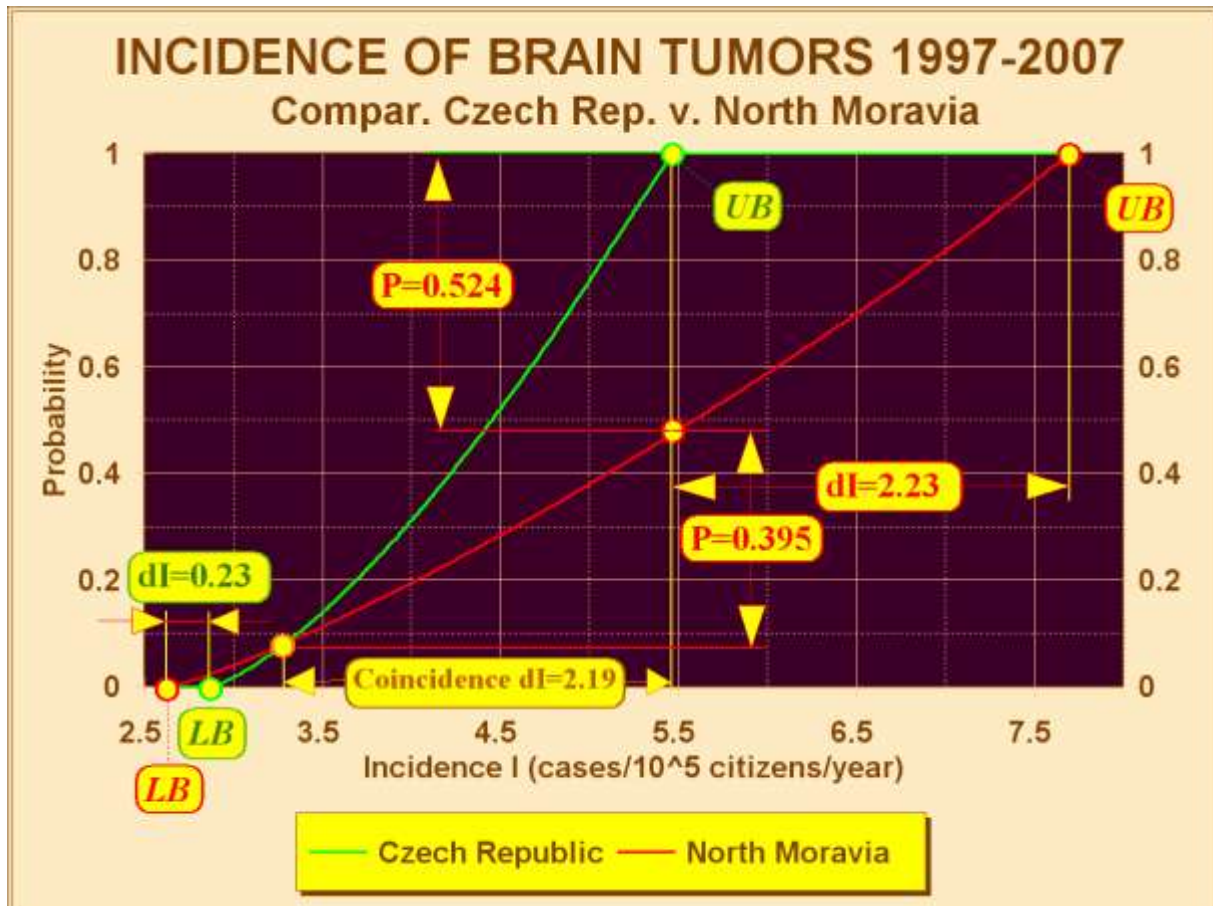


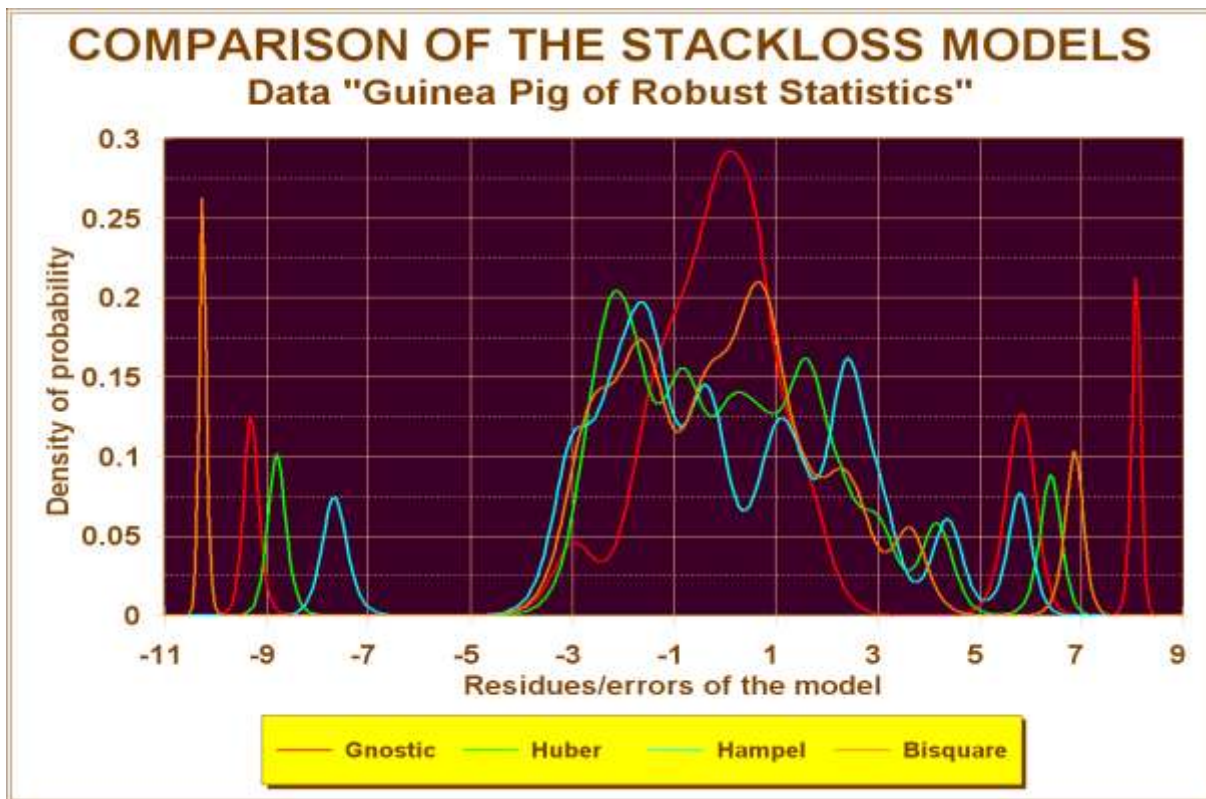
Fig.5: Gnostic test of the hypothesis on brain tumors

The almost sure results are enabled by the finiteness of the distribution functions and by the availability of the estimated values of the bounds *LB* and *UB*.

The origin and way of determination of these figures are shown in Fig.5. These results say more than the (doubtful) statistical statement on a difference between mean values.

### 2.2.3. Identification of the stackloss model

The classical „stackloss“ data alias „the guinea pig of robust statistics“<sup>6</sup> stem from operational data of a plant for the oxidation of ammonia to nitric acid. Three explanatory variables (Air Flow, Water Temperature and Acid Concentration) determine the stack loss in a regression model. Probability density of the model’s errors enable the efficiency of different robust regression models to be compared graphically. The gnostic method (red line) emphasizes the central main cluster of model’s fitting errors about the zero value while clearly separating the outliers.



**Fig.6: Densities of residual errors of four regression models**

Statistics of the compared robust models are in Tab.3 using the same symbols like in Tab.2:

| Method          | R-square | MeanW | MErr   | MAErr | MsqErr |
|-----------------|----------|-------|--------|-------|--------|
| <b>Gnostic</b>  | 0.996    | 0.690 | -0.038 | 0.75  | 1.14   |
| <b>Huber</b>    | 0.940    | 0.968 | -0.027 | 2.03  | 2.71   |
| <b>Hampel</b>   | 0.919    | 0.995 | -0.033 | 2.31  | 2.88   |
| <b>Bisquare</b> | 0.964    | 0.918 | 0.012  | 1.73  | 2.35   |

**Tab.3: Statistical evaluation of the robust models**

<sup>6</sup> Dodge, Y. (1996) The guinea pig of multiple regression. In: *Robust Statistics, Data Analysis, and Computer Intensive Methods; In Honor of Peter Huber's 60th Birthday*, 1996, *Lecture Notes in Statistics* 109, Springer-Verlag, New York.



### 2.2.4. Historical data on fertility in Switzerland

Standardized fertility measure and socio-economic indicators for each of 47 French-speaking provinces of Switzerland at about 1888 are available in R-project's database. Fertility measure is the dependent variable of the robust regression models with four significant explanatory variables expressed in proportions of the population (Agriculture, Education, Catholic and Infant mortality).

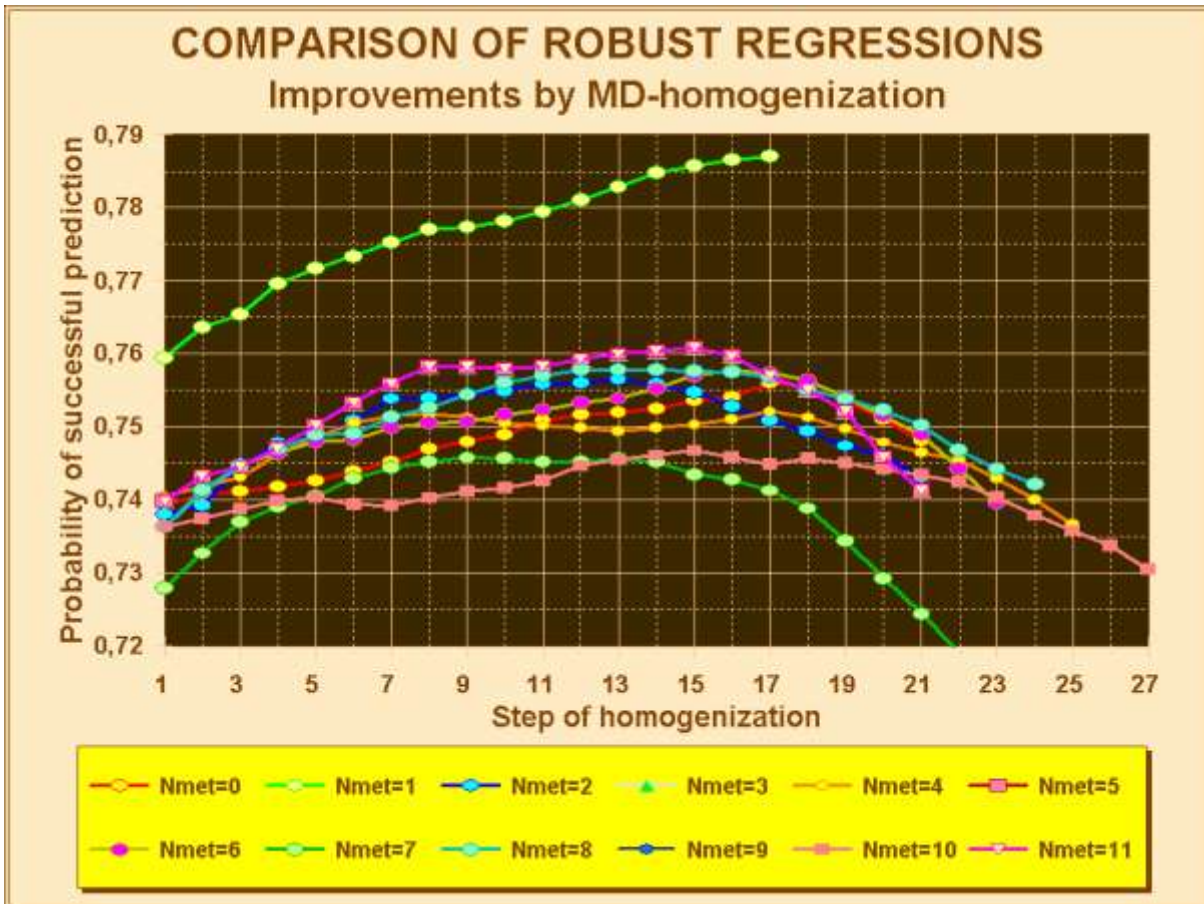
The same characteristics of robust regression methods as in Tab.2 applied to the Swiss data are compared in the Tab.4. They characterize the models' weighted errors resulting by application of the Iterated Weighted Least Square Method (realizing the M-estimators of the robust statistics by the rreg program available in S-PLUS<sup>2</sup> computing environment): The classical method of Ordinary Least Squares (OLS) and ten well-known methods of the robust statistics are compared with the gnostic method realized by the function GWLS.

| Nmet | Method   | R-square | MeanW | MAErr | MsqErr |
|------|----------|----------|-------|-------|--------|
| 0    | OLS      | 0.40     | 1.00  | 7.53  | 9.54   |
| 1    | Gnostic  | 0.98     | 0.46  | 3.36  | 5.33   |
| 2    | Huber    | 0.58     | 0.96  | 6.91  | 8.88   |
| 3    | Hampel   | 0.47     | 0.99  | 7.30  | 9.29   |
| 4    | Bisquare | 0.63     | 0.93  | 6.71  | 8.67   |
| 5    | Andrews  | 0.63     | 0.93  | 6.71  | 8.67   |
| 6    | Fair     | 0.65     | 0.91  | 6.70  | 8.71   |
| 7    | Cauchy   | 0.88     | 0.65  | 5.64  | 7.75   |
| 8    | Logistic | 0.66     | 0.90  | 6.68  | 8.71   |
| 9    | Median   | 0.59     | 1.95  | 3.65  | 6.14   |
| 10   | Talworth | 0.50     | 0.98  | 7.10  | 8.84   |
| 11   | Welsch   | 0.53     | 0.96  | 7.12  | 9.12   |

**Tab.4: Quality comparison of robust statistical regression models with the gnostic one**

### 2.2.5. Stock market predictions

Stock markets evaluate prices of shares using the actually available information, especially the financial statements. For investors a dilemma exists: will the price rise or fall?



**Fig.7: The homogenization/prediction learning process realized by eleven methods**

A robust model can help by predicting the future total return of a considered firm by using its *current working capital*, *total equity*, *turnover* and *earnings* as explanatory variables calculated by using the firm's financial statements. To identify such a model, a homogenized subsample of firms of the whole industry can be used. Homogenization can be achieved by sequentially excluding the firms, the data of which produce outliers in the set of modeling errors. The predictions are gradually improved by eliminating the outliers in a process similar to learning. The homogenization process improving the predictions is depicted in the Fig.7. The numbers of methods (Nmet) are attached to names of methods like in Tab.4. Gnostic method evidently outperforms those of both classical and robust statistics. This result could be supported by the statistical characteristics like in Tab.4.

### 3 The normality problem of uncertain events

#### 3.1 What is to be really normal?

The notion of „normal“ has many interpretations: usual, what you expect, formed or developed in the usual way, average or standard, natural, regular, ordinary, plain, commonplace, perpendicular (in geometry). Statisticians adopted this word for the unique type of (Gaussian) distribution. A problem of a vital importance is connected with this term:



where are the bounds, crossing of which qualifies the change from normal to abnormal?. There are many ways of setting and checking such bounds in different fields: juristic (by laws and norms), technological (by production quality assessment control and automated monitoring and emergency systems), religious and moral (commandments and rules of morality), medical (by confronting the patient's state with the „normal“ one). The problem is that examination of normality is not a primitive<sup>7</sup> operation. Real states and quantities are as well as uncertain as their observations. Boundaries of real events are therefore „fuzzy“. Their determination must cope with uncertainty. A good model of uncertainty is a must.

### 3.2. Statistical setting the bounds of the normal/reference range

Many “standard” distribution functions are defined over an infinite domain. Those having a bound of the domain are ordinarily based on an assumption of an a priori assumed, but not estimated bound. Application of such distributions suffers from double subjectivity – on the distribution's form and on the bound's value. However, neither application of infinite bounds is objective enough.

Consider this simpler case: A given set of random data with a given probability distribution of the set and a given significance level is to be analyzed. The role of the lower bound of the reference range is assigned to the quantile, the probability of which corresponds to the required significance level. The upper bound is the quantile of the probability one minus the significance. A value randomly taken from the set is considered as “belonging” to the reference range if and only if it lies between its bounds.

This approach seriously suffers from at least three instances of subjectivity:

- 1) The probability distribution of a real data sample is rarely available. Its statistical model must then be assumed. The assumption depends on skill of the person doing the analysis. A frequently used assumption to use Gaussian or log-Gaussian distribution is not realistic due to the mismatch of the finiteness of real quantities and the non-finite Gaussian domain. In addition, the zero value acceptable from the log-Gaussian point of view cannot model a realistic, useable observed value. The assumption of “Gaussian normality” is also hidden in some ISO norms requiring the data mean and a  $\pm K$ -multiple of the standard deviation to play the role of the reference bounds. The “normalization” of non-Gaussian distributions by means of data transformation followed by tests of the “normality” of the resulting distribution cannot be a proper remedy because the transformation introduces additional, artificial data interdependencies, which can reduce the reliability of the decisions based on the data.
- 2) Statistical manuals require an *a priori* setting of the significance, normally performed before gathering the data. However, experience shows that if a statistical test does not confirm the expectations or more importantly the requirements of an experimenter or analyst, the change of the “required” significance is frequently applied to make the experiment “successful”. This then leads to the argument that the choice of assumptions is arbitrary and can further degrade the quality of the decision making based on this method.
- 3) Relying on two statistical moments, mean and standard deviation, is also based on the Gaussian assumption. These statistics are efficient in the Gaussian case (sufficient and necessary for estimation of the distribution) but not in cases of

---

<sup>7</sup> In terms of the classical set theory, a primitive piece of knowledge is something “everyone knows”. Example: “everybody knows, if an element belongs to this classical set or not”.



modified distributions. Changes in the distribution form result in changes of probabilities and risks.

Moreover, the non-robust nature of the estimated statistical moments can substantially increase the risk of incorrect decision making.

Unfortunately, this approach is not only very popular on practice but also prescribed by some ISO norms.

### 3.3 Empirical reference range in clinical practice

Success of the quantitative decision making about the normality/abnormality in the health care is critically dependent on the bounds of normality called “reference values” in this field. These are being established by clinical practice mostly in the empirical way. Reference values are defined as bounds of the range of a majority of values obtained by measuring the parameters of a group of “healthy” reference people. The “reference population” is composed from individuals not suffering by the disease, which is to be determined. “Majority” is usually 95%.

This definition is vague because of difficulties of selecting the reference (“healthy”, “normal”) population. The health state of individuals is dependent on many factors, and therefore it is a multidimensional object. The “normality” of the reference population is thus questionable and must be subjected to a reliable test. Application of the reference values to other individuals requires the ability to compare tested individual with the reference population, which is not fully comparable. Moreover, the “cutting of” the five per cent of the reference population is illogical:

- 1) These individuals are originally accepted as “healthy” to be later rejected because of their “peripheral” (“non-healthy”) parameters. Application of the “95%-reference values” to testing the “healthy” individuals can thus cause a false alarm in 5% cases.
- 2) Bounds of the normal range must be broader than the interval between minimum and maximum observed values, because even the extreme “peripheral” but “normal” observed values are uncertain. Therefore, the probability of another “right” observation lying out of this interval cannot be zero.

Determination of reference/normal range of diagnostic parameters should be based on arguments, which would be more rational than the intuitive ones recently being applied.

### 3.4 Gnostic bounds of normality of uncertain events

The importance of the bounds  $LB$  and  $UB$  of the probability distribution has already been documented by the example of the hypothesis test. However, more is to be said on these values:

- 1) They are estimated only from the data without any assumptions on the statistical model of the data, on their independence, on the kind of their distribution function and on significance connected with their values. These values thus represent **objective** estimates entirely independent on the analyst and his wishes and opinion.
- 2) There also exist two other bounds  $LSB$  and  $USB$ , the sample bounds also called the **membership bounds**, because they specify the interval of data values belonging to the **homogeneous** subsample of the data from the interval  $(LB, UB)$  so that – in a general case – the data are belonging to three subintervals:  $(LB, LSB, USB, USB)$ .



The data from the subinterval ( $LSB$ ,  $USB$ ) are members of the homogeneous kernel of the sample.

- 3) The sample bounds  $LSB$  and  $USB$  are also estimated uniquely by using only data, they also are objective characteristics of a data sample.
- 4) The test of homogeneity of a data sample is always performed in estimation of the global distribution function. It is robust. The sample bounds  $LSB$  and  $USB$  are estimable after the data homogenization.
- 5) Homogeneity of a data sample is an important feature supporting the recognition of data “of the same nature” and enabling to separate the homogeneous sample’s kernel from outliers and other data clusters<sup>8</sup>.
- 6) The objectivity and uniqueness of the bounds  $LSB$  and  $USB$  allows to accept them as gnostic normality bounds of a sample of uncertain data and to define the normal data sample as the **homogeneous data sample of data in ( $LSB$ ,  $USB$ )**.

Two examples comparing this concept with the traditional clinic reference ranges follow.

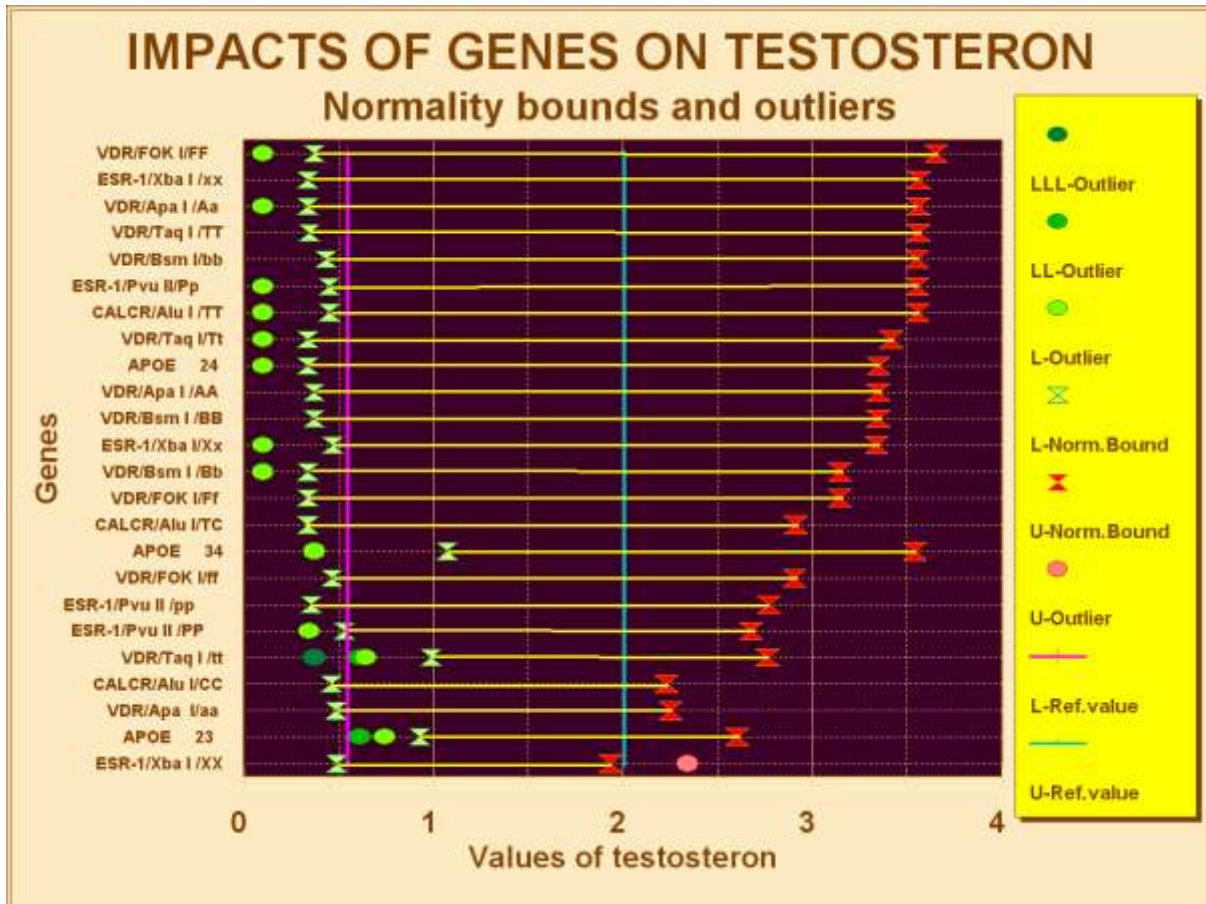
### 3.4.1 Example 1: Normality of testosterone level in postmenopausal women

Hormonal balance in ageing women is an important health factor. Two hormones are of special importance, testosterone and estradiol. Like other hormones, the bounds of their reference range are established and accepted by clinical practice. To test the normality bounds, values of testosterone and estradiol measured in 114 postmenopausal women were analyzed<sup>9</sup> along with other hormones and with eight genes, each of which had three alleles. The results on testosterone are in Fig.8. All patients were sorted into 24 subgroups with respect to their genetic factor. Estimation of the bounds  $LB$  and  $UB$  then followed for each of the groups.

---

<sup>8</sup> This prevents analyst from mixing up “pears with apples” or “men with women” and likewise.

<sup>9</sup> Data were made available thanks to courtesy of prof. Ivana Žofková, DrSc., Charles University and Endocrinological Institute of the Czech Academy of Sciences, Prague.



**Fig.8: Gnostic bounds of normality of testosterone level depending on genetic factors**

The magenta and cyan vertical lines indicate the generally accepted clinical reference values in Fig.8. The green and red X-marks are the normality bounds *LSB* and *USB*. The elliptic marks show the outliers having the abnormal values. It can be seen that the true upper bound of the normal values could be even about double of the „traditional“ maximum value. The „usual“ reference values were confirmed by the gnostic ones only in one from 24 cases. The lower bounds of the range could also be not supported by gnostic results, they also could cause mistaken conclusions on the patient’s health. This means, that hormonal diagnosis and resulting remedy actions could be not adequate to the patient’s state and that a more careful identification of the homogeneous subgroups of the patients taking in account genetic information was necessary.

### 3.4.2 Example 2: Normality of estradiol level in postmenopausal women

Results related to estradiol are presented in Fig.9. Number of cases substantially exceeding the upper bound of the reference range was lower, but the maximum values reached were much larger exceeding the classical reference value up to four times. Discrepancies were also revealed in the lower bounds: the gnostic genetically influenced bounds were found deep below the clinical minimum in eleven from 24 cases.

The clinical reference bounds were 0.04 and 0.15. Availability of data enabled the values of the estradiol’s mean and standard deviation to be estimated and the log-normal distribution



to be confronted with the reference range. The symmetric confidence interval from  $P=0.025$  through  $P=0.975$  was applied to adhere to clinical practice. The lower quantile of the log-normal distribution appeared to be  $Q(P=0.025) = 0.0399$ , i.e. it nearly coincided with the clinical bound. Unlike this, the upper quantile of this distribution was  $Q(P=0.975) = 0.476$ , i.e. it exceeded the clinical upper bound more than three times. This can be interpreted as a contradiction between the statistical and clinical ways of estimation of the reference ranges.

Anyway, the reasonability of the gnostic approach and its applicability to solution of this and similar important problems have been confirmed.

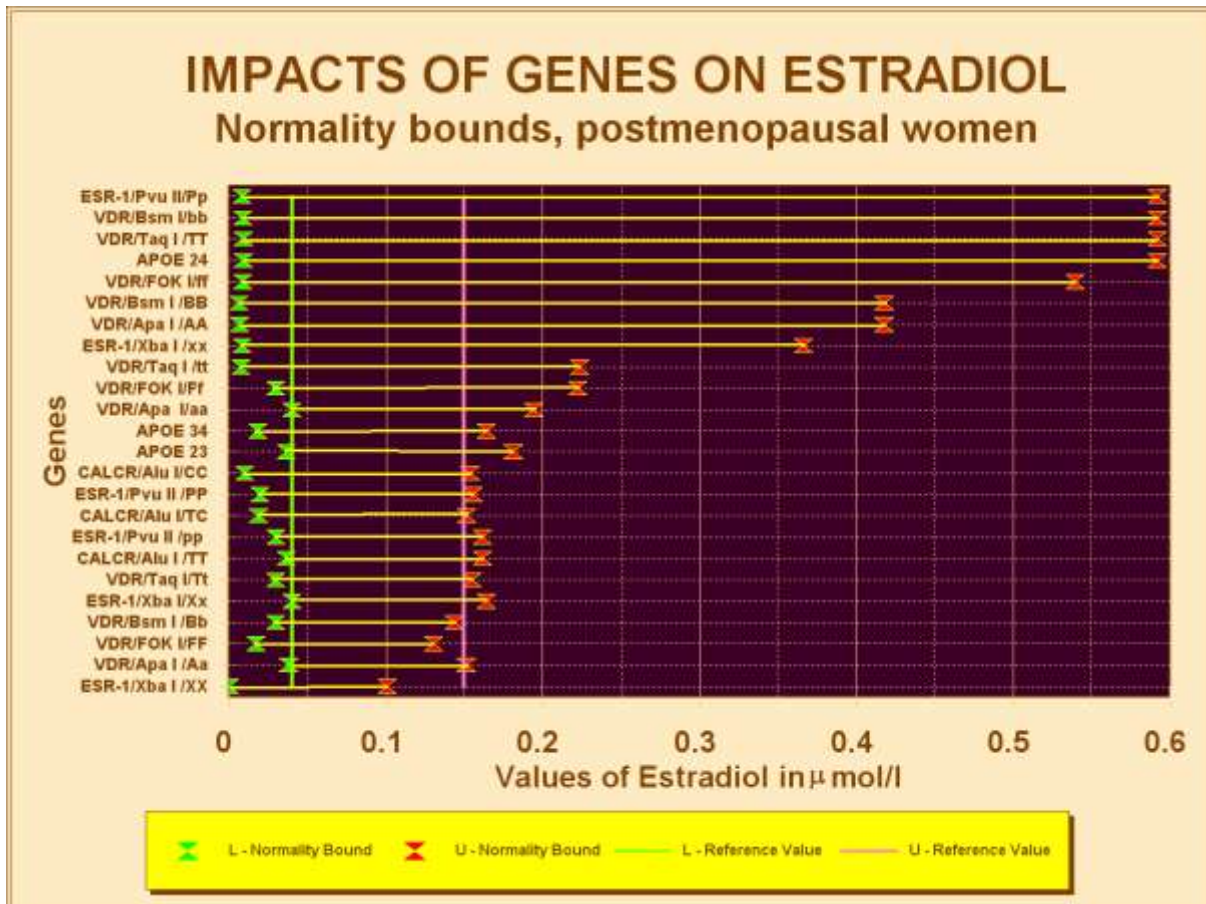


Fig.9: Gnostic bounds of normality of estradiol level depending on genetic factors

#### 4. Conclusions

Theoretical backgrounds of statistics and mathematical gnostics are different, although their goals are similar: to treat the uncertain data as best as possible. Statistics has been originally developed to be applied to treatment especially of mass data contaminated with not very strong uncertainty. The assumption of a weak uncertainty (formulated in the Central Limit Theorem as mathematical existence of the mean and variance) enabled many statistical linear/quadratic models operating in Euclidean spaces to be created and applied. Development of computers lead to emphasizing the problem of robustness and to creating of robust statistical models, many of which exist in curved spaces with many kinds of artificially imputed metrics. Unlike this, mathematical gnostics is a theory of individual uncertain data



and small data samples not assuming some uncertainty limitations. It generally operates in curved spaces endowed by Riemannian metrics determined not by an analyst but objectively by the treated data. Robustness of methods is their objective and natural feature. However, as shown in theory, the gnostic nonlinear characteristics of uncertainty and of its impact on estimates converge to linear/quadratic statistics when the data errors are sufficiently small. Using this circumstance, one can come to the point of view, that models of mathematical gnostics are generalizations or extensions of the statistical ones legally applicable to small data sample of strongly uncertain data.

Comparison of results of parallel applications of methods of mathematical gnostics and methods of both classical and robust statistics presented by this report supports the statement that the new approach deserves to be used and further developed in favor of the both theory and practice.