



---

## 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**

---

### **– Definition of case studies for demonstrating the relevance of innovative methods for data treatment –**

Due date of delivery: *31/07/2008*

Actual submission date: *06/10/2008*

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

Duration: *48 Months*

Lead contractor organisation name for this deliverable: *IPH*

Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)	
<b>Dissemination Level</b>	
<b>PP</b>	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** Definition of case studies for demonstrating the relevance of innovative methods for data treatment  
**ID** D1.5\_v3.doc  
**Revision** Version 2  
**Revision Date** 04/08/2008  
**Author** P. KOVANIC/IPH and T. OCELKA / IPH

## Approvals

	Name	Company	Date	Visa
<b>Author</b>	P. KOVANIC	IPH	04/08/2008	P. Kovanic
<b>Co-Author</b>	T.OCELKA	IPH	04/08/2008	T.Ocelka
<b>WP Leader</b>	A. MARCOMINI	UNIVE	27/09/2008	A. Marcomini
<b>Coordinator</b>	F. BOIS	INERIS	27/09/2008	F. Bois

---

## Document history

Revision	Date	Modification	Author
0	18/01/2008	Template made available to the WP leaders	F. Bois
1	04/08/2008	First version	P. Kovanic
2	24/09/2009	Revision after comments by WP1 and WP2 leaders	P. Kovanic



## **Contents**

<b>INTRODUCTION.....</b>	<b>4</b>
<b>BRIEF DESCRIPTION OF THE CASE STUDIES TO BE ELABORATED.....</b>	<b>4</b>
<b>BRIEF DESCRIPTION OF THE DATABASE TO BE APPLIED TO THE CASE STUDIES .....</b>	<b>6</b>
<b>INTENDED TASKS TO BE CARRIED OUT IN CASE STUDIES .....</b>	<b>7</b>
<b>METHODS TO BE APPLIED.....</b>	<b>9</b>
<b>CONCLUSIONS .....</b>	<b>11</b>



## INTRODUCTION

The goal of the Objective 3 of the 2FUN project was to improve and develop uncertainty models for further health management. The state-of-the-art of approaches to this problem was characterized as not taking into account in risk assessment the uncertainties caused by a high variability in time and space of data measured in monitoring emissions and concentrations of pollutants. To optimise the information contained in input data and incorporate sources of uncertainty in exposure assessment, it was decided to include into the project the task of reviewing, testing and providing guidelines of robust methods for improving the quality of environmental input data, to fill data gaps and thus to optimise the 'upstream' data treatment<sup>1</sup>. Requirement of robustness of the analytical methods was motivated by the experience, that classical methods of mathematical statistics behave not satisfactorily, when applied to data not obeying a priori assumptions to statistical models of uncertainty. Moreover, neither modern methods of robust statistics did not prove universally successful in applications to data measured under the hard conditions of environmental monitoring. This motivated orientation to alternative methods, among which the methods of mathematical Gnostics attract attention due to its theoretically and practically justified applicability to small samples of strongly dispersed data.

Methods of mathematical Gnostic will be applied to a specific case study with real environmental health data to demonstrate the efficiency of this approach. The description of the selected dataset and the presentation of the planned applications of the selected methods are illustrated in the following chapters.

## BRIEF DESCRIPTION OF THE CASE STUDIES TO BE ELABORATED

### ***CASE STUDY I.: ROBUST TREATMENT OF DATA MEASURED BELOW THE SENSITIVITY TRESHOLD***

Measurement of parameters of the environment as concentrations of emissions in the air and/or pollutants in water is extremely difficult because of the broad range of the quantities to be measured. The norms establish their tolerated maxima, which must be kept very low, because the real danger of their effects results from their permanence and long-term accumulation in living organisms. The lower bound of the range is very low and appears to be in many cases below the sensitivity threshold of the measuring technology called the Limit of Detection (LOD). Measuring usually realizes repeatedly in different locations and in several time moments. Information on the state of the environment is contained in results of measurements as a whole, all data bear information. It would not be reasonable to neglect data measured below the LOD, because these are the "best" ones, the most desirable from the point of view of population safety. Omission of these data could significantly overestimate the real danger.

However, data below the LOD (the *left-censored* data) represent only a part of the problem. Strong volatility of environmental data along with emergency situations resulting from accidents in industry and/or transport result in extremely high concentrations of dangerous substances, which sometimes overcome the upper range of measurability. These *right-censored* data along with outliers and the measurement range covering many orders of magnitude contribute to difficulty of the monitoring. Neither these data cannot be neglected, because they signal a real or suspected emergency. Both left- and right-censored data must be thus taken in account, and information they contain is to be extracted and used by the data treatment methods.

Ways of estimation of these data became therefore the object of intensive investigation. A detailed review of methods developed and used in handling the left-censored data is in Bacarelli at al. (2005)<sup>2</sup>. Following methods are discussed in this paper:

- 1) Deletion.
- 2) Simple substitutions consisting in imputing following values instead of the left-censored data: zero,  $LOD/2$ ,  $LOD/\sqrt{2}$ ,  $LOD$ .

---

<sup>1</sup> 2FUN-project Description of Work

<sup>2</sup> A.Bacarelli at al.: Handling of dioxin measurement data in the presence of non-detectable values: Overview of available methods and their application in the Seveso chloracne study, Chemosphere 60 (2005) 898-906



### 3) Distribution-based methods.

These methods can be applied independently on the kind of the measured quantity. However, importance of this task also forces the investigators to looking for an approach applicable to an important special case. An example of this can be found in Needham et al. (2007)<sup>3</sup>. The approach is based on observation, that a roughly constant ratio of concentrations of two certain congeners can exist. This can be used for estimation, when one congener is measurable while the other's value are left-censored.

All these approaches can be subjected to criticism: the deletion of censored data as well as the imputed substitutions do not use the data information efficiently. Probability distribution methods also violate the data by imputing them an a priori assumed distribution. The log-normal distribution frequently applied is unnatural by fundamental reasons: its domain is infinite, while real measured quantities are always bounded. Relying on some special features of the measured variables not only limits the applicability of the method, but also leads to doubts on universality of the special observation.

Generally speaking, the idea of using the probability distribution as an inherent and natural regularity ruling the observed process can be considered reasonable for everyone believing in non-random character of the natural processes and in their subjection to regularities. The probability distribution represents such a regularity and data contain information of the distribution. If so, then the "small" (left-censored data) could be subjected to the same distribution as the "large" ones. But to exploit this idea requires a realistic estimate of the probability distribution to be applied, because an a priori assumed distribution function can be far from reality.

Mathematical Gnostics<sup>4</sup> developed a distribution estimating method, which proved itself in applications to problems from various fields. It is therefore worth trying this approach to real left-censored data and to test its efficiency. This alternative methodology can be applied within the 2FUN project thanks to the decades lasting long-term development making available the Gnostic software recently realized using the S-PLUS<sup>5</sup> mathematical and statistical environment.

## ***CASE STUDY II.: ROBUST TESTING OF HYPOTHESES***

Environmental research along with routine monitoring of the parameters of the environment frequently raises a question, which is unavoidable, because it must be answered to initiate an action. Decision making in management of health and environmental control is conditioned by fast and reliable answering the questions frequently formulated as decision problem of accepting or rejecting a ("zero") hypothesis against a competitive ("alternative") hypothesis. Danger potentially present in consequences of the decision making, makes it necessary not only to choose the decision, but also to evaluate risks connected with the decision. There has been decision making methodology developed in statistics, the basic scheme of which includes following steps:

- 1) Define zero hypothesis to be tested and the alternate hypothesis.
- 2) Set significance level  $P_s$  of the test.
- 3) Prepare probability distribution function  $P_o(q)$  of the zero hypothesis, where  $q$  is the quantile (numeric representation of the independent variable).
- 4) Prepare probability distribution function  $P_a(q)$  of the alternate hypothesis.
- 5) Find quantile  $Q_s$  such that  $P_o(Q_s) = P_s$ .
- 6) Find probability  $P_a(Q_s)$ .
- 7) Determine the risks connected with the decision:
  - I. Error alpha =  $P_s$  (risk of the 1-st kind) resulting from the rejection of the true zero hypothesis.
  - II. Error beta =  $1 - P_a(Q_s)$  (risk of the 2-nd kind) resulting from rejection of the alternate hypothesis, which is true.

---

<sup>3</sup> Needham L.L. et al.: Assigning concentration values for dioxin and furan congeners in human serum when measurements are below limits of detection: An observational approach. *Chemosphere* 67 (2007) 439-447

<sup>4</sup> Kovanic P.: Gnostic Theory of Uncertain Data, DrSc. Thesis, The Institute of Information Theory and Automation of Czechoslovak Academy of Sciences, Prague, (in Czech), 152 pp. (1990)

<sup>5</sup> S-PLUS is a registered trademark of Insightful Corp., Seattle, Washington, USA



Both risks are functions of the required significance level. Its value controls the ratio of the both risks in dependence on given distribution functions.

Stumbling block of all the statistical tests including this one is in trustfulness of the distribution functions. Many applications are based on a priori assumed distribution functions. There exist natural processes, for which reliable mathematical models of distributions were theoretically developed and successfully applied. Examples are known especially from physics: exponential distribution of the radioactive decay of nuclei, Maxwell distribution of velocities of gas molecules and several others. However, as already mentioned, universally applicable mathematical models of parameters of the environment do not exist. Instead, experimentally estimated distribution functions are to be applied. Quality of these estimates determines actual risks of hypotheses testing and dangers resulting from the decisions.

The problem of setting the level of significance and resulting power of the test (equaling 1-beta) deserves a comment. Significance level usually considered as acceptable in statistics is 0.02 or 0.05. But real life can prepare situations requiring other reliability of outcome of decisions: Passengers of an airplane would not be glad to know, that an accident is probable with one of twenty flights (with probability 0.05). On the other hand, it is known, that about a half of marriages results in divorces. This can be interpreted so, that probability 0.5 of a success is acceptable for many people deciding to marriage. In health management as well as in environmental control analogous situations occur, where decision making can reasonably accept significance level differing from the “usually” required statistical values. So, for example, lowering the contamination level of the environment to one halve can surely justify some investments and efforts. This does not restraint importance of hypotheses testing. On the contrary, it emphasizes the requirement of applicability of the test not only to very low significance, but to decisions based on tests of an arbitrary (even not given a priori) power. This can be warranted only by using trustful estimate of probability distributions based on real data.

## **BRIEF DESCRIPTION OF THE DATABASE TO BE APPLIED TO THE CASE STUDIES**

### ***ORIGIN OF THE DATABASE***

There exists a database suitable for both intended case studies<sup>6</sup>. 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 factors that influenced the dioxin levels were investigated on the basis of a questionnaire. The survey was realized in collaboration of three Czech institutions:

- 1) National Institute of Public Health, Prague,
- 2) The 3<sup>rd</sup> Faculty of Medicine, Charles University, Prague,
- 3) The Institute of Public Health, Ostrava.

The aim of the study 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.

### ***JUSTIFICATION OF THE CHOICE OF THE DATABASE***

Decision to choose the described database can be supported by the following reasoning:

- 1) Having already made standard statistical analysis available offers a chance

---

<sup>6</sup> Černá, M. et al., Levels of PCDDs, PCDFs, and PCBs in the blood of the of the non-occupationally exposed residents living in the vicinity of a chemical plant in the Czech Republic, *Chemosphere*, Vol 67, Issue 9, 238-246(2007),  
doi:10.1016/j.chemosphere.2006.05.104



- a) to demonstrate, that using alternative methods can yield more information than the classical approach,
  - b) to compare the quality and completeness of results obtained by the standard and alternative approach.
- 2) Uniqueness of this database is based on its thoroughness manifested by inclusion of measurements not only of 29 congeners of organic pollutants found in blood of the exposed individuals, but also of potentially important personal identification of the survey's participants:
- i) Location of their settlements relative to the chemical plant.
  - ii) Age.
  - iii) Gender.
  - iv) Two indicators of the individual's health state (subjective/objective).
  - v) Alcohol consumption.
  - vi) Body-mass-index (BMI).
  - vii) Smoking activity.
  - viii) Details on consumption of locally produced food.
  - ix) Other personal details.

The database enables thus posing a broad range of questions on mutual dependences between parameters of contamination and their effects observable with population. Examples of resulting tasks intended for treatment in both studies follow.

## **INTENDED TASKS TO BE CARRIED OUT IN CASE STUDIES**

### ***TASKS FOR THE CASE STUDY I***

The Case Study I. should be opened by brief statements of the Gnostic theory, on which the Gnostic programs will be based enabling to perform following actions:

- 1) To estimate robust probability functions of selected individual pollutants (namely of TCDD and PeCDD relevant for judgment on the method of Needham at al.) and of the sum of all 29 pollutants measured in the blood of 80 persons. All available data (both censored and uncensored) are to be used, while applying the Gnostic method of making use of all the data.
- 2) To estimate distribution functions of TCDD (whose 63 concentrations from eighty appeared to be below the LOD) obtained by imputation of values used in methods of the simple substitution into the LCD (Low Censored Data items):
  - a)  $LCD = 0$
  - b)  $LCD = LOD$
  - c)  $LCD = LOD/2$
  - d)  $LCD = LOD/\sqrt{2}$
  - e) Using only uncensored data ( $LCD > LOD$ ).
- 3) To robustly estimate (by using the distribution functions) the first and second statistical moments (means and standard deviations) of all the distributions to compare the estimation errors obtained by the standard methods with the alternative (Gnostic) one.
- 4) To compare the distribution function of the TCDD obtained according to the Needham's idea as  $0.4 \times PeCDD$  with the actual TCDD distribution estimated by the Gnostic method using all data.
- 5) Using the distribution functions, estimate the ratio of concentrations TCDD/PeCDD resulting from the Czech database.
- 6) To support the confidence in reliability of the Gnostic results, perform the following experiment:



- a) To be sure about the true value of the results obtainable in handling the censored data by the Gnostic method, estimate the distribution function of a data measured properly (with no data censored). Particularly, the distribution of the sum of concentrations of 29 pollutants is suitable for this purpose, because it does not comprise any censored data.
- b) Order the data of this sample in the ascending manner.
- c) Take the 10-th, 20-th, 30-th, 40-th, 50-th, 60-th and 70-th member of the ordered sample as values  $LOD_k$  ( $k=1, \dots, 7$ ) of the Limits of Detection.
- d) Form for each value of “k” a data sample by substituting the least 10k values of the uncensored sample instead of the original (uncensored) values. Other data (the  $(10k+1)$ -th through the 80-th) leave unchanged.
- e) Estimate the distribution functions of the seven “left-censored” sample containing  $10k/80$  censored data and  $(80-10k)/80$  true data values.
- f) Compare the distribution functions of censored data with the true distribution function of all 80 data.
- g) Evaluate the errors of the method numerically by the following statistics: 1Q, 2Q, 3Q (three quartiles), mode (quantile of the maximum density), the first and second statistical moments obtained by numerical integration of the distribution function (mean value and standard deviation).

### ***TASKS FOR THE CASE STUDY II.***

Using the database defined above, estimate the necessary distribution functions, and test following hypotheses:

- 1) Summary amount of organic pollutants in blood depends on the distance between the person’s settlement and the considered chemical plant.
- 2) Amount of organic pollutants in blood is an increasing function of the individual’s age allowing thus to introduce the indicator called the “rate of accumulation”.
- 3) The rate of accumulation depends on the distance of the person’s settlement from the considered chemical plant.
- 4) The rate of accumulation depends on the kind of pollutant.
- 5) The rate of accumulation depends on the person’s gender.
- 6) The amount of accumulated organic pollutants negatively affects the health state.
- 7) There is a non-negligible impact of the MBI on the rate of accumulation.
- 8) Smoking does not help in keeping the MBI low.
- 9) Consuming of small amounts of alcohol decreases the amount of accumulated organic pollutants.
- 10) Consuming of small amounts of alcohol improves the person’s health state.
- 11) Consuming of food locally produced in vicinity of the chemical plant affects negatively the health of inhabitants of the region.
- 12) Living far from a chemical plant (but in Czech Republic) does not protect population from contamination by organic pollutants.

These hypotheses may be true as well as wrong. The tests should not only decide between the zero hypotheses and their alternatives, but also estimate risks alpha and beta. Moreover, robust estimates of the finite domains of probability distributions should allow to derive some statements valid almost surely (events, to which probability zero or one is attached).



## METHODS TO BE APPLIED

### *METHODS OF GNOSTIC 1D-ANALYSIS*

One-dimensional Gnostic analysis (marginal, or 1D-analysis) is performed by the software package called GMASWX, where X is the number of the actualized version. It consists of the 80 functions written in S-language of the S-PLUS<sup>7</sup> mathematical-statistical package. Many programs written using this language can be run by using the noncommercial package of the R-project. However, differences in functions ensuring the multidimensional optimization (necessary even for estimation of the Gnostic distribution functions) made it necessary to have two versions of the GMASWX, one for the S-PLUS and the other for the R-project.

The most important and complex function of the GMASWX is the function called GNDF. This function along with other functions of the GMASWX package performs estimation of four versions of Gnostic distribution functions:

- A) EGDF...Estimating Global Distribution Function,
- B) ELDF... Estimating Local Distribution Function,
- C) QGDF...Quantifying Global Distribution Function,
- D) QLDF... Quantifying Local Distribution Function.

Estimating functions differ from the Quantifying ones by the opposite robustness, the Estimating ones being robust with respect to outlying data, while Quantifying distributions prefer the peripheral data being robust with respect to inner disturbances of data samples. The Global distributions characterize the sample's distribution as a whole, assuming (and robustly testing) its homogeneity manifested by unimodality of the density of probability. This feature along with the robustness to peripheral data allows the EGDF to be applied for robust estimation of the bounds of the data support (function's domain) and of the scale parameter directly from data. These bounds along with the scale parameter and the sample's data completely define the distribution function without any assumptions on statistical model of the data. Unlike EGDF, the ELDF is universally applicable even to the inhomogeneous data samples, because of its unlimited flexibility determined by the chosen scale parameter. This makes the ELDF suitable for depicting multimodal density functions and to the marginal (onedimensional) cluster analysis decomposing an inhomogeneous data sample into several homogeneous ones.

The Quantifying versions of the distribution functions are usable, when the inner robustness is required. Such tasks include, for instance, the quality assessment control. In such applications, the "ordinary", "usual", "proper" values of the monitored variable must not initiate any warning or emergency signals while the values reaching extreme values represent the real danger, which must activate the emergency or quality control system. In such cases the role of the "preferred" values and the "noise" are exchanged with respect to the case of robustness to outliers.

All the four distribution functions can be applied under various conditions arising from requirements of the tasks and from data features. The modifications are achieved by the GNDF's arguments allowing following applications:

- a) to both additive and multiplicative data,
- b) to data samples, the lower and/or upper bound of which is a priori determined by a constraint of fundamental nature,
- c) to homoscedastic as well as heteroscedastic data (a constant or a variable scale parameter/variance),
- d) to censored data of three kinds: left-censored, right-censored and interval data,
- e) to data having some a priori known (not uniform) weights,
- f) to data, some of which have repeated values (such as of the form of a histogram). These can be "compressed" by the GNDF to rationalize the computing.

The applications covered by the functions of the GMASWX include making use of the distribution functions and results of the Gnostic theory to many tasks, not only the "standard" ones like the estimating of the probability and its density to given quantiles and estimating the quantiles to given probability. From "non-usual" applications, the following can be mentioned:

---

<sup>7</sup> S-PLUS is the registered trade-mark of the Insightful Corp., Seattle, Wa., U.S.A.



- 1) Robust filtering of data.
- 2) Robust and unique estimates of the bounds of the “membership interval”, significance of the membership being not given arbitrarily but uniquely determined by the data. (This function answers the question “Does a data item  $x$  belong to the data sample  $X$ ?”, or “Is the  $x$  an element of  $X$ ?”. There also is a third formulation: “Is  $x$  an outlier with respect to sample  $X$ ?”).
- 3) Robust testing of the data sample’s homogeneity.
- 4) Two methods of homogenization of inhomogeneous data samples, i.e. of extracting a homogeneous subsample from a sample.
- 5) Robust estimating of the a posteriori weights of individual data dependent on the relation of each observed data value to the estimate of its true value.
- 6) Robust monitoring, filtering and prediction of probability of reaching some given values of a time series.
- 7) Comparison of distribution functions and their densities.
- 8) Robust (probability distributions based) estimation of covariances and correlations.
- 9) Robust estimation and classification of typical subintervals of the data domain.
- 10) Robust testing of hypotheses.

To solve tasks of both case studies, not only some of these functions, but also the functions of the multidimensional analysis will be applied.

### ***METHODS OF GNOSTIC MD-ANALYSIS***

Gnostic multidimensional (MD-) analysis is based on the theoretically derived non-linear measures of errors and data weights enabling the information resulting from the solution of the regression problem to be maximized. Gnostic software for this analysis exists as a package called GMDAY, where Y is the number of the version. It consists of 34 functions written in the S-PLUS language, which can be run not only within the S-PLUS environment, but also by using the noncommercial R-project.

The main tasks of the package are carried out by three programs of robust MD-regression applicable to data, to logarithms of strictly positive data and to data probabilities instead of data. All three programs can be applied either to explicit or implicit regression analysis.

The explicit regression models correspond to the usual types, which approximate the dependent variable by a set of functions of explanatory variables. Such a model is realistic under the assumption, that there exists one exceptional variable, which is dependent on the explanatory variables without affecting them. Such a variable may not always exist in real systems. Take a human body’s measurable variables, like blood pressure, pulse rate, temperature, local electric potentials and others. It is obvious, that all of them interact, while none of them can be considered as the “only dependent” one, not affecting the others. In spite of this, possibility to approximate the interdependences of the variables by a mathematical model cannot be denied. In the implicit models, the regularity of interactions of all variables “having equal rights” is expressed by an equation system, where the “explanatory” side is formed by combination of functions of all variables, while the role of the “dependent” variable is played by a constant.

There exists an approach in robust statistical theory called the “M-estimate”, which can be applied to the regression modeling as iteratively reweighted least square method (WLSQ). This method applies a weight to each equation of the system iteratively in dependence on the actual equation error. The weighting (“influence”) function results from some theoretical principles. The S-PLUS system offers a choice between nine functions developed by the experts of robust statistical theory. There exist Gnostic versions of the weighting functions derived by optimization of information or other Gnostic measures of errors<sup>8</sup>. One of these has been included in the Gnostic programs along with the robust statistical versions. This allows efficiency of all versions in application to real data to be compared.

---

<sup>8</sup> Kovanic P.: A New Theoretical and Algorithmic Basis for Estimation, Identification and Control, Automatica, Vol.22, No.6, 657-674 (1986)



The main task of the regression modeling is the robust estimation of the parameters of the model. However, high quality of the results of the approach enabled several unusual but useful task to be developed and realized in the form of functions of the package GMDAY:

- 1) Monitoring an MD-time series of parameters of an MD object enabling robustly, but reliably derived warning and emergency signal to be generated.
- 2) Robust estimation of an MD-model of a set of MD objects resulting in multidimensional ordering of the objects evaluating the individual relations of each object to the “collective” or cross-sectional model.
- 3) Robust multidimensional cluster analysis enabling a non-homogenous MD data sample to be decomposed into a set of homogeneous data subsamples.
- 4) Robust prediction of an MD-object.

Both packages GMASWX and GMDAY are completed with various functions needed for preparing data to analysis and for depicting the results.

## CONCLUSIONS

One of the main objectives of the 2-FUN project was formulated in the project’s Description of Work by the following: “A main breakthrough in the tools proposed by 2-FUN lies with the explicit accounting of uncertainty throughout the full mechanistic computation chain and with the possibility to functionally integrate different data classes, seamlessly crossing different time and spatial scales. All these are essential components of a user-friendly yet scientifically robust information system“. The information flow necessary for decision making and management of the environmental health field is unavoidably mediated by the handling and interpreting the data obtained by difficult measurements, realized under hard conditions and constraints. Experience has shown, that classical statistical analysis does not meet the specific requirements of the field. This motivated looking for suitable alternative methods to fight the strong uncertainty met with monitoring the parameters of the environment and the damage caused by its contamination.

There are many alternatives to statistical methods recently available, but none of them did not prove to be the universally tool suitable to satisfy specific needs and requirements of the environmental control. As a promising candidate suitable for the 2-FUN project has been proposed the approach based on the Gnostic theory of individual uncertain data and small data samples. Its applicability and efficiency was demonstrated by many examples presented in the 2-FUN report<sup>9</sup>. To verify the suitability of the Gnostic methodology and to demonstrate it in more detail, two case studies have been planned in 2-FUN project oriented to two difficult problems of the environmental control and of analysis and interpretation of the environmental health data:

- I.) Robust treatment of data measured below the sensitivity threshold.
- II.) Robust testing of hypotheses.

Tasks to be carried out by both studies were defined in detail. A database originated in a large Czech health survey (oriented both to levels of contamination by organic pollutants, and to factors possibly connected with impacts of contamination) has been selected as a source for both studies. Two software packages of algorithms intended for the studies were shortly described.

The necessary elements of the intended studies were thus defined and made available.

---

<sup>9</sup> P. Kovanic and P. Ciffroy: Review and Tests of Methods for Robust Data Treatment, Deliverable D1.4 of the 2-FUN project (2008), 22 pp.