New Look at Analytical Data Through the Gnostical Method

Tomáš Paukert* and Ivan Rubeška
Czech Geological Survey, Malostranské náměstí 19, 118 21 Prague 1, Czechoslovakia
Pavel Kovanic
Institute of Information Theory and Automation, Czechoslovak Academy of Science, Pod vodárenskou věží 4, 180 00 Prague 8, Czechoslovakia

The gnostical theory represents a new, powerful approach towards evaluating data files. This paper describes the application of a gnostical analyser which may help in finding outliers, testing the homogeneity of sets of data and classifying individual data. As an example, its use for ascertaining the recommended values for reference materials is demonstrated by means of homogeneous and heterogeneous sets of data.

Keywords: *Robust estimator; gnostical method; recommended value; reference material*

New analytical procedures may best be verified by analysing certified reference materials (CRMs) with well established concentration values for the constituents to be determined. Apart from this, CRMs are also important for quality assurance and quality control in analytical laboratories. The determination of the true analyte concentration in a CRM from experimental data is therefore an important task toward which the efforts of many workers have been directed.

The list of papers dealing with this problem and using statistical methods is fairly extensive. The data to be treated, however, do not always represent a homogenous population. In such instances, robust estimators less susceptible to non-homogeneities in experimental data have been applied by many workers,[1–3] but the outcome has often been dubious, if not frustrating.[4,5] In this paper we describe the application of a new non-statistical 'gnostical' method.

The gnostical theory (GT) is an axiomatic–deductive mathematical theory. It has been developed as an alternative to mathematical statistics for the treatment of data containing uncertainty, having a statistical model that is not known, or data with a statistical characterization that does not adequately describe the essence of phenomena. The GT may also be successfully applied to data for which, for various reasons, a limited number of observations or measurements are available, *i.e.*, the quality of information is poor or the data are influenced by the contribution of a rare, but strong disturbance of an undefined character.

During the development of GT programming, five different classes of gnostical programs have been distinguished. One of them, the gnostical analysis discussed in this work, may be applied for in-depth analysis of limited data files, robust estimations of expectancy or probability of phenomena, distribution functions of data files and their density, robust cluster analyses, investigation of file homogeneity and of the equivalence or diversity of two or more files, robust estimations of location parameters and of scale parameters of individual clusters and classification of individual data according to the degree of their relevance to individual clusters.

Gnostical Theory of Uncertain Data

Sound data deserve to be given greater weight than unsound data. However, two problems exist in this connection: (*i*) how to distinguish the unsound data from the sound data; and (*ii*) how to optimize the weights to make maximum use of the information.

Heuristic approaches cannot ensure either optimization or the universal applicability required by practice. It is well known that 'the most practical tool is a good theory'. Unfortunately, a good theory of such a complex problem cannot be both simple and directly acceptable by way of some 'common sense' considerations. In contrast to other theories involving uncertainty, the GT of data files is based on an axiomatic theory of an individual uncertain datum and on a data composition axiom. The axioms of this theory have a simple algebraic nature. To illustrate the functions of the GT, first the main equations will be described.

Consider the ith real-valued A_i (an 'additive' datum) together with its 'multiplicative' equivalent,

$$Z_i = \exp(A_i) \tag{1}$$

having a strictly positive value. For a positive real scale parameter s, a real variable $z > 0$ and for a sample of N data, define the auxiliary quantities

$$q_i(z,s) = (Z_i/z)^{2/s} \tag{2}$$

for use in the calculation of N 'fidelities':

$$f_i(z,s) = 2/[1/q_i(z,s) + q_i(z,s)] \tag{3}$$

and 'irrelevancies':

$$h_i(z,s) = [1/q_i(z,s) - q_i(z,s)]/[1/q_i(z,s) + q_i(z,s)] \tag{4}$$

Within the framework of the GT, irrelevance plays the role of the distance between z and Z, the fidelity being the weight of the datum Z_i. Introduce the arithmetic mean

$$f(z,s) = \sum_{i=1}^{N} f_i(z,s)/N \tag{5}$$

of the fidelities and define the symbol $h(z,s)$ for the irrelevances analogously. Let $w(z,s)$ be the function of weight defined by the relationship

$$W^2(z,s) = [f(z,s)]^2 + [h(z,s)]^2 \tag{6}$$

The distribution function generated by the individual datum Z_i is then

$$L_i(z,s) = [1 + h_i(z,s)]/2 \tag{7}$$

having the density

$$l_i(z,s) = \frac{\mathrm{d}[L_i(z,s)]}{\mathrm{d}z} = f_i^2(z,s)/(z,s) \tag{8}$$

At least two theoretical results of the GT are immediately applicable to the data files provided by analytical chemistry, *i.e.*, the two types of data distribution functions (DDF), global (GDF) and local (LDF). These functions play a role analogous to the probability distribution functions. The field of application of the gnostical distribution functions is, however, much broader, as these functions do not rely on some statistical

* To whom correspondence should be addressed.

assumptions or probabilistic concepts. They characterize the data patterns and expectations of the subject deduced from their particular shape. For a weak influence of the uncertainty of the data, *i.e.*, minor data errors, the two gnostical DDFs differ to only a negligible extent. However, their behaviour may be different for major data errors. This is a desirable feature that makes these gnostical DDFs very useful for analyses of totally different types of data files.

The LDF $L(z,s)$ is simply the arithmetic mean of the distribution functions [eqn. (7)] of individual data:

$$L(z,s) = \sum_{i=1}^{N} l_i(z,s)/N \qquad (9)$$

The GDF $G(z,s)$ is obtained using the function of weight w [eqn. (6)]:

$$G(z,s) = \sum_{i=1}^{N} l_i(z,s)/w(z,s) \qquad (10)$$

The LDF is a relatively universal instrument which can be applied even to data files containing non-homogeneities such as individual subclusters. The derivative of the LDF, called the data density function, has a multi-modal form, in which each mode corresponds to a subcluster. Being a monotonous function for an arbitrary data file, the LDF can always be determined. In the special case of the reasonability of a statistical interpretation of data, the LDF [eqn. (9)] can be an asymptotically consistent kernel estimate of the Parzen type[6] of probability distribution function. In such a case, the GT is used as a source of a theoretically justified kernel which generates remarkably clear and smooth density curves, even for small data files. In the much more general case of data files that do not allow statistical interpretation, the equation of the DDF is still valid as a continuous model of the distribution of the expectation that a 'new' datum of the same nature as the 'old' data will have a certain value. The LDF is 'locally robust' in the sense that its local form, corresponding to a subinterval of a data range, does not influence its form in another subinterval. This is due to the steep descent of the gnostical kernel [eqn. (8)].

Unlike the LDF, the GDF has theoretical justification only for special data files of homogeneous type. Such files should have a unimodal data density function. For a non-homogeneous data file, the GDF may lose the fundamental feature of a distribution function, its monotony.[7] This fact makes it possible to perform an efficient test of the homogeneity of the data files. The limited flexibility of the GDF permits the estimation of the proper scale parameter which characterizes the spread of the data. Most unique, however, is the GDF, which is globally robust in the sense of the low sensitivity of its shape with respect to the 'outliers' and also to all the other 'peripheral' subclusters of the data. This leads to a highly reliable prognosis of rare events (of values of the GDF for very small or very large quantiles). Such tasks often appear in practice in connection with random quality controls, studies of lifetimes, *etc.* This type of gnostical distribution function has no known statistical analogy.

The LDF and GDF differ substantially in their dependence on the scale parameter s. Let $F(N)$ be the 'empirical' distribution function of the data file. The function $F(N)$ has the known form of an irregular staircase. The LDF of the same sample can be made to approach the $F(N)$ as close as required, choosing a sufficiently small positive value for the s parameter. In contrast, the maximum distance of the GDF has a minimum for a 'best' s, which can be recognized as a robust estimate of the scale parameter s. Hence the GDF is as close as possible to the $F(N)$. The choice of s determines the resolution power of individual clusters of data files.

An overall survey of the GT, with a detailed description of the mathematics involved, can be found elsewhere.[8]

Application of the GT in Analytical Chemistry

In routine practice, the evaluation of data is usually carried out by statistical calculations. The accuracy of an analytical procedure may best be verified by analysing CRMs with well established 'recommended' values (RV). For rock CRMs, the RVs are usually derived from round-robin tests with the participation of many laboratories. If the central values (location parameters P_1) for particular elements derived by statistical evaluation of the data are mostly concordant, the assignment job is relatively 'easy' and the RV may be established from the various central values. If the P_1 are discordant, a decision has to be made as to the suitability of the analytical methods employed for the particular concentration levels. Some elements in the Periodic Table are notoriously troublesome for quantitative determination, which may be demonstrated by repeated analyses. Differences of as much as 100% or more are sometimes encountered, especially when determining element contents at ultra-trace levels.

To demonstrate the possibilities of the new non-statistical mathematical method discussed, we used data from the 1987 Compilation Report on the Ailsa Craig Granite, AC-E.[9] The AC-E reference material, prepared with great care, was distributed to 128 laboratories in 29 countries. From the submitted data, RVs for many elements were successfully established. Some elements were determined by a limited number of laboratories only and for some elements discrepancies in the results were evident. Two elements were selected as examples to demonstrate the potential of the GT for evaluation of results: first, the results for europium, which represent good concordance of reported data, and second, results for cobalt, which have a significantly heterogeneous file of data.

Homogeneous Data File

The results reported for europium are a good example of a homogeneous data file. This element was determined in 41 laboratories. All the results are given in Table 1.

Table 1 Results for Eu (ppm). For explanation of abbreviations used, see text

Laboratory	Value	Procedure	Laboratory	Value	Procedure	Laboratory	Value	Procedure	Laboratory	Value	Procedure
1	1.39	EMN	11	1.9	EMN	21	2	CSP	31	2.07	EMN
2	1.4	CSP	12	1.92	CSP	22	2	CSP	32	2.09	CSP
3	1.6	CSP	13	1.92	EMN	23	2	CSP	33	2.1	CSP
4	1.7	CSP	14	1.93	CSM	24	2	CSP	34	2.1	EMN
5	1.8	CSP	15	1.94	CSP	25	2	CSP	35	2.16	EMN
6	1.8	EMN	16	1.94	EMN	26	2	EMN	36	2.23	EMN
7	1.86	EMN	17	1.94	EMN	27	2	EMN	37	2.3	BSM
8	1.87	EMN	18	1.96	EMN	28	2.03	EMN	38	2.3	CSP
9	1.9	CSP	19	1.98	EMN	29	2.04	EMN	39	2.4	CSP
10	1.9	EMN	20	1.99	EMN	30	2.05	ASM	40	2.68	EMN
									41	3.1	ASM

Each method is designated by a three-letter code, the first letter indicating the method of sample preparation and the last two the method of determination.[9] For sample preparation, A = acid decomposition, B = fusion with fluxes, C = dissolution + separation, D = mixture with buffers and E = simple physical conditioning; for determination, AA = atomic absorption spectrometry (AAS), FX = X-ray fluorescence spectrometry (XRF), SM = mass spectrometry, MN = nuclear methods, SF = flame photometry and SP = direct reading atomic emission spectrometry (AES).

For the first evaluation of data files the data were treated by the GT using the GDF. The resulting distribution function, presented in Fig. 1, may help in investigating the file homogeneity. This type of estimate is based on the *a priori* assumption that the tested data file is homogeneous. It is robust with respect to peripheral data (outliers). In Fig. 1 both the distribution function and the density function of the data file are shown. The dotted lines indicate the degree for goodness of fit of the distribution function estimated by the Kolgomorov–Smirnov test with plotted intervals for 10, 20, 50, 95 and 99% probability. The central value for europium corresponding to the maximum of the curve was calculated to be 2.007 ppm.

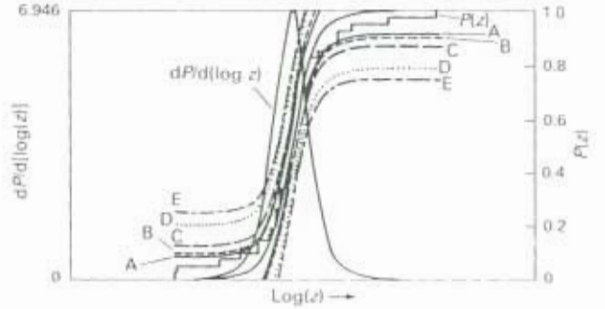

Fig. 1 Global distribution of the Eu data file. z, Concentration of analyte; $P(z)$, distribution function of the concentration z and $dP/d(\log z)$, density function of the concentration z. Kolgomorov-Smirnov test: A, 10; B, 20; C, 50; D, 95; and E 99%

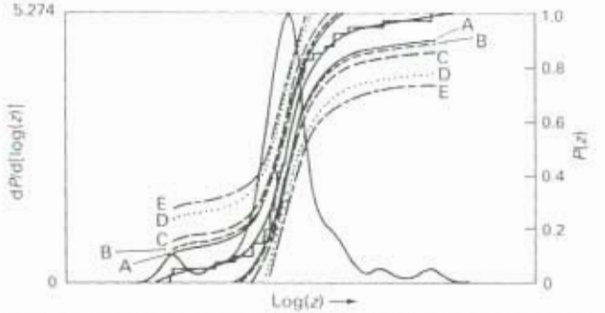

Fig. 2 Local distribution of the Eu data set. A, 10; B, 20; C, 50; D, 95; and E, 99%

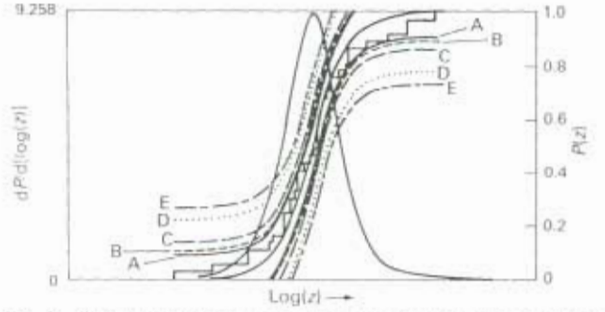

Fig. 3 Global distribution of the Eu data file after restriction of 4 data. A, 10; B, 20; C, 50; D, 95; and E, 99%

Even though the peak in Fig. 1 appears clear, the distribution curve was recalculated using the LDF model, which should distinguish local clusters in a data file. The result, presented in Fig. 2, shows three small maxima, obviously corresponding to the values 1.39, 1.4, 2.68 and 3.1 ppm. These may be interpreted as outliers. After their exclusion, the calculation of the global distribution was repeated and the result is presented in Fig. 3. The central value calculated according to the LDF model is 1.980 ± 0.109 ppm. After having eliminated the outliers, it decreases to 1.970 ppm. The value of 0.109 ppm is the standard deviation derived from the GT calculations.

In Table 2 this P_L is compared with the various mathematical parameters reported in the AC-E compilation.[9]

The robustness of the GT-derived P_L may be demonstrated by the following example. The x_a and the gnostical central values (GCVs) were first calculated from the entire number of 41 analyses. Then, the one-step restriction was executed; this means that the values 2.68 and 3.1 ppm were trimmed off and x_a and the GCVs were calculated again from the remaining 39 values. The comparison of the calculated results may be seen in Table 3. The GCV evidently changes considerably less than the arithmetic mean.

Heterogeneous Data File

We selected a data set with non-homogeneous results, at the same concentration level as europium and with a similar number of analyses. The 40-value file for cobalt from the AC-E CRM[9] fulfilled these requirements well. The cobalt values have a great spread, ranging from 0.07 to 10 ppm (see Table 4).

The outcome of the calculation of the global distribution from these data is presented in Fig. 4. The broad maximum clearly reflects the wide range of the results reported. As the congruence between the GDF and the empirical distribution function is poor, the data file cannot be considered as homogeneous. The data set was recalculated once more, using the LDF model, and the result is shown in Fig. 5. On the local distribution curve, three distinct maxima appear, the location of which, as calculated by the GT, results in three different concentration levels for cobalt, namely 0.16; 1.39 and 4.9 ppm.

In Table 5 the central values, as reported for the AC-E in the compilation[9] but split according to the method of analysis used, are given. A comparison with the results given by GT immediately suggests that the three maxima correspond to central values of the results by nuclear methods only, by optical spectrometric methods (AAS + AES) and by XRF spectrometry, respectively. It is worth mentioning that the RV chosen was 0.2 ppm and was based on the nuclear methods set of data.

Table 2 Mathematical parameters calculated from 41 results for Eu[a]

Parameter	N	x_a	M	MG	x_p	x_{geo}	x_{cm}	x_g	GCV
Value	41	2	2	1.99	2	1.99	2	1.99	1.98

[a] The derived RV = 2 ppm.[9] N = number of analyses; M = median; x_p = preferred mean calculated after $1 \pm s$ elimination; x_g = gamma central value; x_a = arithmetic mean; MG = Gastwirth median; x_{geo} = geometric mean; x_{cm} = dominant cluster mode; GCV = gnostical central value. For more information about the statistical parameters, see ref. 9.

Table 3 Comparison of robustness (for definitions of parameters see Table 2)

Number of results	x_a	GCV
41	2.1017	1.983
39	1.962	1.974

Table 4 Results for Co (ppm). For explanation of abbreviations used, see text

Laboratory	Value	Procedure	Laboratory	Value	Procedure	Laboratory	Value	Procedure	Laboratory	Value	Procedure
1	0.07	EMN	11	0.21	EMN	21	1.65	EMN	31	4.9	DFX
2	0.091	EMN	12	0.28	EMN	22	2	AAA	32	5	EFX
3	0.118	EMN	13	0.5	AAA	23	2	ASP	33	6	ASP
4	0.13	EMN	14	0.6	EFX	24	2	ASP	34	6	BFX
5	0.14	EMN	15	1	AAA	25	2	BAA	35	6	EFX
6	0.153	EMN	16	1	ASP	26	2	EFX	36	7	DFX
7	0.18	ASP	17	1	EFX	27	3	AAA	37	7	EFX
8	0.2	EMN	18	1.15	EMN	28	3	EFX	38	9.5	EFX
9	0.2	EMN	19	1.4	AAA	29	4	EFX	39	10	AAA
10	0.21	EMN	20	1.4	EMN	30	4.7	DFX	40	10	EFX

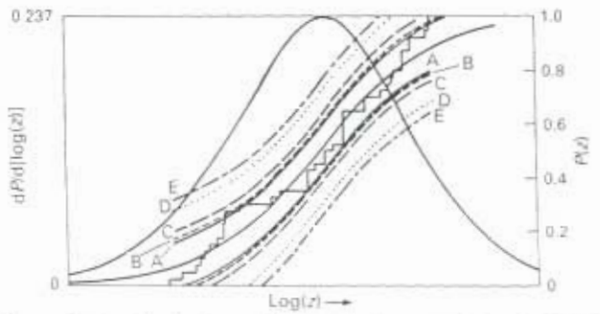

Fig. 4 Global distribution of the Co data file. A, 10; B, 20; C, 50; D, 95; and E, 99%

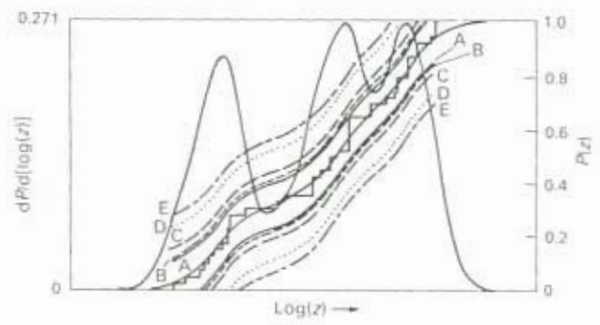

Fig. 5 Local distribution of Co data file. A, 10; B, 20; C, 50; D, 95; and E, 99%

Table 5 Mathematical parameters for the analyses of Co. For explanation of abbreviations used, see text

Calculated parameter	Method used				
	MN	AA	SP	FX	Total
N	14	7	5	14	40
x_a	0.43	2.84	2.24	5.05	2.69
M	0.2	2	2	4.95	1.53
MG	0.19	1.82	1.7	4.98	1.69
x_p	0.16	1.65	1.3	5.29	1.45
x_{geo}	0.24	1.88	1.34	4.01	1.15
x_{cm}	0.19	1.38	—	4.87	0.17
x_g	0.19	1.73	1.61	4.96	1.4
GCV	0.16–1.39–4.90				

From the analytical point of view, it seems surprising that for a frequently determined element such as cobalt, such large discrepancies in analytical data may occur. One should nevertheless bear in mind that the AC-E was specially prepared as a CRM for the rare earth elements and the analysis results for cobalt are only a useful by-product of no particular interest. As 0.2 ppm is an unusually low cobalt content in rocks, it is probably well below the concentration range for which the instruments are routinely calibrated. The readings by optical spectrometry and XRF were plausibly evaluated simply by extrapolation to lower concentration. The greater error by XRF probably reflects the fact that 0.2 ppm of cobalt is closer to the limit of detection by XRF than by optical spectrometry.

However, the most remarkable outcome, in our view, is that the gnostical analyser applied to the entire set of data did provide three different results which, in addition, are in fairly good concordance with the statistical analysis applied separately to the results by different analytical methods. This demonstrates the possibilities of the GT when applied to non-homogeneous sets of data in discerning separate data files.

Conclusions

We have tried to draw attention to a new, powerful tool for treating analytical data provided by the gnostical theory. This is demonstrated by applications of the gnostical analyser to the data from a collaborative study[9] and for deriving recommended values for the CRM rock AC-E. Although the GT cannot provide RVs from insufficient data, it can reveal their 'heterogeneity'. For such sets the GT is highly sensitive and may distinguish the independent files even without any additional information. With respect to P_L calculation, it also exhibits high robustness, thus providing a theoretically based, practical substitute for empirically derived robust estimators. The GT is still being developed and its applications are very promising. The program system entitled 'interactive analyzer GA2' has been adapted for use on an IBM PC. Readers interested in the GT should contact P. Kovanic.

References

1 Ellis, P. J., and Steele, T. W., *Geostand. Newsl.*, 1982, **2**, 207.
2 Lister, B., *Geostand. Newsl.*, 1984, **7**, 171.
3 Abbey, S., *Geostand. Newsl.*, 1988, **9**, 241.
4 Abbey, S., paper presented at Geoanalysis 90, Huntsville, Canada, 1990.
5 Abbey, S., *Chem. Geol.*, 1992, **95**, 123.
6 Parzen, E., *Ann. Math. Stat.*, 1962, **35**, 1065.
7 Baran, R. H., *Automatica*, 1988, **24**, 283.
8 Kovanic, P., *Automatica*, 1986, **22**, 657.
9 Govindaraju, K., *Geostand. Newsl.*, 1987, **11**, 203.

Paper 2/02281H
Received May 1, 1992
Accepted August 5, 1992