130

CZECHOSLOVAK ACADEMY OF SCIENCES

# CZECHOSLOVAK JOURNAL OF PHYSICS

REPRINT

VOLUME B 36 1986

# GNOSTICAL ANALYSIS OF INTERNATIONAL ACTIVITIES IN PHYSICS

P. Kovanic

*Institute of Information Theory and Automation, Czechoslovak Academy of Sciences*
*Pod vodárenskou věží 4, 182 08 Praha 8, Czechoslovakia*

J. Vlachý

*Kaňkovského 1241, 182 00 Praha 8, Czechoslovakia*

Data on country and subfield distributions of the world publication output and on national citation records in physics are evaluated by the means of a new, gnostical theory. Substantive and robust results suggest possible wider use of this approach in scientometrics and research management.

## METHOD

Gnostical theory [1—3] derives from two simple axioms a mathematical model of data acquisition disturbed by uncertainty, statistical model of which is unknown or even unjustifiable. The theory is based on laws governing the uncertainty of each individual datum such as variational principles of virtual kinematics of real data and of their dynamics closely related to both entropy and information of data. The knowledge of an ideal gnostical cycle for each real datum enables to develop algorithms for optimal handling of data. Software based on gnostical theory maximizes the information obtained and yields data characteristics robustness of which is optimal with respect to outlying or inlying data. Applications include the robust estimation of both location and scale parameters of small data samples and of their generalized correlations, cluster analysis, estimation of probability and a nonparametric estimation of probability distribution, nonlinear discrete filtering, prediction and smoothing, identification of systems under strong disturbances and adaptive setting of alarm systems, robust identification of regression models, etc. The main advantage are algorithms efficient in applications to small samples of bad data.

## ANALYSIS

An $i$-th uncertain observation $z_i$ of an exact ("true") quantity $z_0$ will be supposed to have the form

$$z_i = z_0 \exp(s\Omega_i) \tag{1}$$

being a member of a data sample $Z := \{z_1, \dots, z_n\}$. The quantities $z_0$ and $s$ will be called the scale parameter and the location parameter of the sample $Z$, respectively. The quantity $\Omega_i$ parametrizes the influence of uncertainty on the datum $z_i$. Only $z_i$ is known in (1), the other parameters are to be estimated within the analysis. It can be shown [2] that the gnostical variance of the outlier-robust type has the form

$$v = 1 - \frac{1}{n}\sum_i^n f_i^2 \tag{2}$$

where the quantity $f_i$ (the "fidelity" of the $i$-th datum) is

$$f_i = 2/((z_i/z_0)^{2/s} + (z_0/z_i)^{2/s})\,. \tag{3}$$

Table 1

Assessment of international publication patterns by ratios between major physics subfields.

a) Statistical and gnostical characteristics of data samples (data are taken from [7, 5]).

b) Tolerance intervals of location parameters and typicality intervals.

ELEM — The physics of elementary particles and fields, NUCL — Nuclear physics, CON6 — Condensed matter: structure, thermal and mechanical properties, CON7 — Condensed matter: electronic structure, electrical magnetic, and optical properties, MATE — Materials science.

| No | ratio of physics subfields | categ. of count. | numb. of count. | a) | | | | b) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | characteristics | | | | tolerance interval | | interval of typical values | |
| | | | | statistical | | gnostical | | | | | |
| | | | | $\bar{z}$ | $\sigma$ | $\tilde{z}_0$ | $\tilde{s}$ | $z_L$ | $z_U$ | $z_{TL}$ | $z_{TU}$ |
| 1 | ELEM + NUCL | I | 27 | 0·49 | 0·25 | 0·55 | 0·92 | 0·53 | 0·57 | 0·39 | 0·78 |
| | CON6 + CON7 + MATE | II | 23 | 0·45 | 0·37 | 0·36 | 0·89 | 0·34 | 0·38 | 0·26 | 0·50 |
| 2 | ELEM | I | 25 | 0·52 | 0·40 | 0·39 | 0·90 | 0·37 | 0·40 | 0·28 | 0·53 |
| | NUCL | II | 18 | 1·00 | 0·90 | 0·74 | 1·24 | 0·46 | 1·27 | 0·31 | 1·91 |
| 3 | CON7 | I | 26 | 1·30 | 0·43 | 1·36 | 0·61 | 1·26 | 1·32 | 1·04 | 1·62 |
| | CON6 | II | 21 | 2·95 | 3·40 | 1·52 | 1·27 | 1·45 | 1·59 | 0·96 | 2·41 |
| 4 | CON7 | I | 26 | 0·85 | 0·35 | 0·87 | 0·66 | 0·86 | 0·88 | 0·69 | 1·10 |
| | CON6 + MATE | II | 22 | 1·56 | 1·74 | 0·80 | 1·24 | 0·76 | 0·85 | 0·51 | 1·27 |
| 5 | CON6 | I | 26 | 2·05 | 1·18 | 1·83 | 0·80 | 1·77 | 1·88 | 1·36 | 2·44 |
| | MATE | II | 21 | 2·07 | 2·48 | 0·84 | 1·28 | 0·77 | 0·93 | 0·50 | 1·42 |

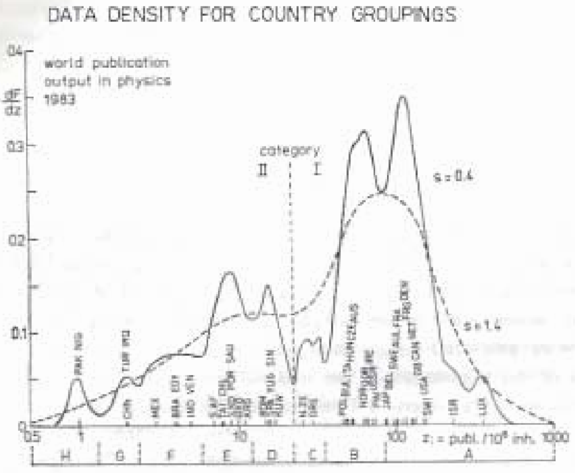

Fig. 1

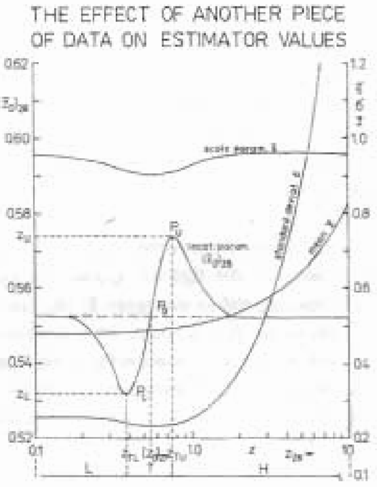

Fig. 2

The gnostical variance characterizes the spread of the data of the sample $Z$ but it plays also another important role: The mean square of fidelities (equalling to $1 - v$) is a good estimate of the data density of the sample $Z$. It can be interpreted as a gnostical estimate of the probability density of the unknown variable $z_0$. For each $z_0$ (given a scale parameter $s$) the data density $d = 1 - v$ can be easily evaluated by (2) and (3).

## RESULTS

The procedure has been first tested on a distribution of 1983 international publication output (50 countries) in physics [4, 5]. In fig. 1 the full line represents the data density for $s = 1{\cdot}4$. This value has been gnostically estimated from an entropy conservation concept. By variating the scale parameter a more detailed insight into clustering of individual countries is feasible. When choosing $s = 0{\cdot}4$, minima on the dotted line reveal seven groups of countries (when forgetting about an "outlier" LUX). A remarkable gap at about 22·5 publications per million inhabitants separates the key groups A and B from the less important C — H. All countries A — C are highly developed economically [6] and it is obvious that they are also highly developed in physics when assessed scientometrically (in fact, there is a strong correlation between the measures of national publication output and per capita GNP [4]). The separation into groups A and B corresponds roughly to the division East—West (cf. disciplinary clusters [5]).

Each of the groups has two important parameters: the location of the maximum and the value of the maximal density. The former quantity can be taken as an estimate $\tilde{z}_0$ of the unknown quantity $z_0$. It will characterize the location of a particular cluster of the data and can be determined analytically as

$$\tilde{z}_0 = \arg\max_{z_0} (d) \tag{4}$$

which is given by a solution of the nonlinear equation

$$\tilde{z}_0 = \sqrt[4]{\left[\left(\sum_i^n \tilde{f}_i^3 z_i^2\right) / \left(\sum_i^n \tilde{f}_i^3 z_i^{-2}\right)\right]} \tag{5}$$

where $\tilde{f}_i$ is obtained from (3) by substituting $\tilde{z}_0$ instead of $z_0$. This equation has a solution for each of the clusters contained in the data sample $Z$. As seen from (5), such a location parameter is highly robust with respect to the data lying out of the immediate neighbourhood of the $\tilde{z}_0$. In case of unimodal density of data and very weak uncertainties this location parameter approaches the value of the arithmetic mean of the data but otherwise it principially differs from the mean by robustness and multiplicity. It is therefore better to characterize data clusters by this location parameter than by the arithmetic mean.

To demonstrate this we take as another example relative weights of physics publication output by subfields in the same 50 countries (data filed for [7] and presented graphically in [5] on p.165, left side of the figure). Countries belonging to groups A — C (more than 22·5 publ./$10^6$ inh.) are considered as category I, the less productive countries D — H fall in category II. Parameters of location $\tilde{z}_0$ together with arithmetic means $\bar{z}$ are given in tab. 1a for all ten data samples. Standard deviations $\sigma$ and gnostical scale parameters $\tilde{s}$ are given as well. In cases when the density curve attains two maxima the location of the maximum of the main cluster has been included in the table (second maxima appeared in the cases 3-I and 5-I because of the extremal low values for LUX).

The arithmetic mean would not be adequate here as the data are obviously very far from being normally distributed and large amounts of negative publication rates would have to be expected. Means in the table are strongly influenced by rare values of large ratio: in seven out of ten cases they exceed the robust location parameters $\tilde{z}_0$. From fig. 2 it is apparent what happens when the

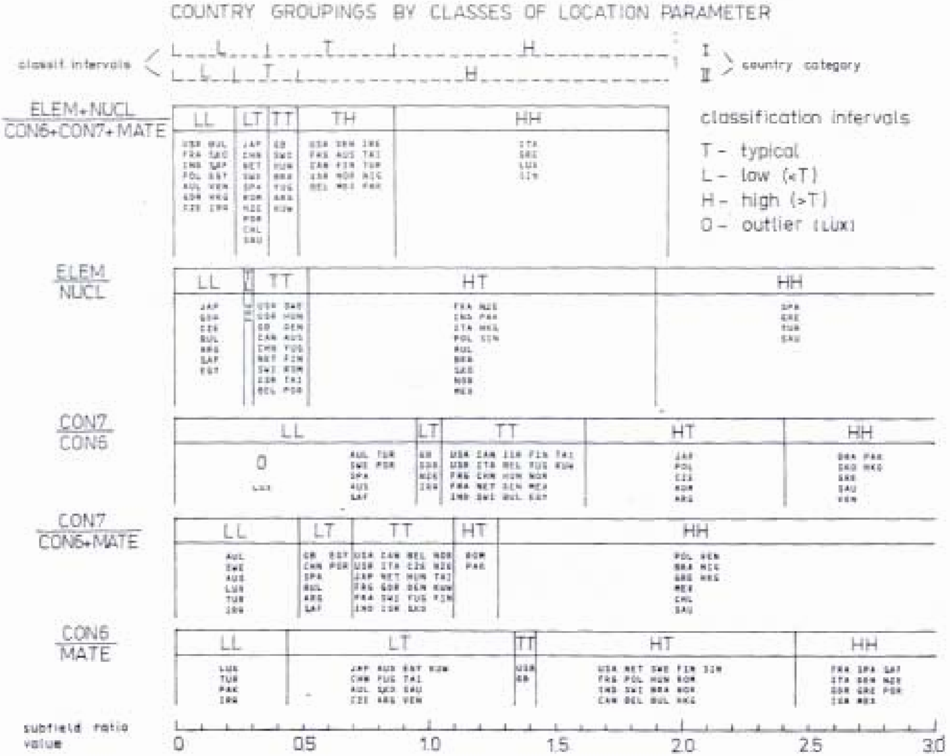

Fig. 3

sample 1-I is extended by inclusion of another datum $z_{28}$. The horizontal axis corresponds to this datum, the vertical axis is used for the location parameter $(\bar{z}_0)_{28}$ of the extended sample. Both statistical measures $(\bar{z}_0)_{28}$ and $(\sigma)_{28}$ grow without limits as the value of $z_{28}$ increases: one bad datum can destroy them notwithstanding the 27 good data in the sample. The asymmetry presents another problem — the statistical parameters react quite differently to small and large values of $\bar{z}_{28}$. But the studied ratio may have both extremal values and it is no reason why to treat them differently. Usage of some robust statistical estimators (median, etc.) may not be a universal remedy [8].

The behaviour of gnostical characteristics $(\tilde{z}_0)_{28}$ and $\tilde{s}$ manifests high robustness with respect to changes of $z_{28}$. The curve for $(\tilde{z}_0)_{28}$ provides two important points $P_{\mathrm{L}}$ and $P_{\mathrm{U}}$ with coordinates $(z_{\mathrm{TL}}, z_{\mathrm{L}})$ and $(z_{\mathrm{TU}}, z_{\mathrm{U}})$, respectively. Location parameter lies between two bounds $z_{\mathrm{L}}$ and $z_{\mathrm{U}}$ for all $z_{28}$, the interval $(z_{\mathrm{L}}, z_{\mathrm{U}})$ can thus be interpreted as a "tolerance interval" of the location parameter. The quantities $z_{\mathrm{TL}}$ and $z_{\mathrm{TU}}$ divide the interval $(0, \infty)$ into three subintervals $L$, $T$ and $H$. For all $z_{28}$ from the interval $T$ ("typical" values) growing $z_{28}$ increases the location parameter $(z_0)_{28}$ while the other values $z_{28} < z_{\mathrm{TL}}$ ("low") and $z_{28} > z_{\mathrm{TU}}$ ("high") cause an opposite, "unnatural" effect (see tab. 1b).

There of course exist many statistical methods of robust estimation of both the location and scale parameters. Effectiveness of several ones has been compared on series of classical data (Short 1763, Newcomb 1882, Michelson 1879) [8] but gnostical estimators have been shown to be superior over all of them [9].

Results from tab. 1b may be used for further analysis: First, in all five cases the differences between categories I and II are significant in the sense that there exist no common points of their toleration intervals of location parameters. Second, each datum on subfield proportions can be adjoined to an interval $L$, $T$ or $H$ from two points of view using bounds determined for category I or II. The outcome of such a classification is summarized in fig. 3.

DATA DENSITIES FOR COUNTRY GROUPINGS

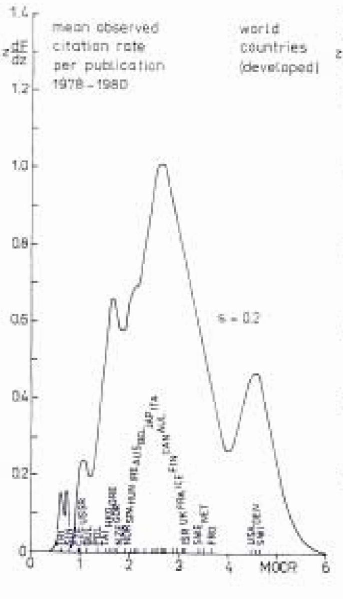

Fig. 4.

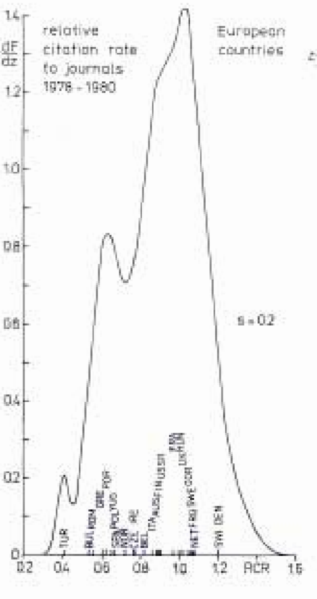

Fig. 5.

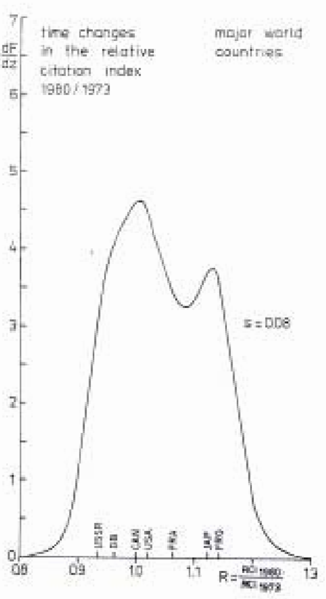

Fig. 6.

Three other recent studies on physics communication patterns have been probed with the same basic gnostical tool. Data density for a set of 32 developed countries (our category I) ranked by mean observed citation rate per physics publication [10] is plotted in fig. 4. For scale parameter $s = 0{\cdot}2$, five groups of countries can be distinguished, providing a quality pattern fairly different from the output pattern in fig. 1. The countries are grouped around maxima located at the values of MOCR = 4·6, 2·6, 1·6, 1·1 and 0·7. On the other hand, 25 European countries, characterized by their relative citation rate in physics (relative contribution to the quality of journals used, i.e. mean citation rate per article divided by mean citation rate of the journals used) [11] concentrate in two or three groups when applying the same $s = 0{\cdot}2$ in fig. 5, at RCR = 1·0 and 0·6, without much clue as to e.g. their per capita national wealth. Finally, 1973—1980 changes in the relative citation index (the ratio of percent of citations received by that country to percent of publications from that country) as obtained from seven oustanding national physics communities [12] can be pictured again more plastically in fig. 6 than by merely comparing the growth rates. For $s = 0{\cdot}08$, FRG and JAP indeed excell at $\mathrm{RCI}_{1980}/\mathrm{RCI}_{1973} = 1{\cdot}3$, while the actual core of countries of course surrounds the maximum at 1·00 (the world standard).

All three last cases would deserve a more substantive discussion on, e.g. how the quality (citation) image of international physics activities does or does not coincide with the output (publication) image [13], or on what is specific in this approach against the former distribution [7], scatter [4] and cluster [5] studies of world physics literature.

## CONCLUSION

Evidence from several case studies have illustrated interesting possibilities of improving the assessment of scientometric data by non-statistical methods derived from the gnostical theory.

Received 27. 9. 1985.

### *References*

[1] Kovanic P.: Probl. Contr. Inform. Theory *13* (1984) 259—274.
[2] Kovanic P.: Probl. Contr. Inform. Theory *13* (1984) 303—319.
[3] Kovanic P.: Probl. Contr. Inform. Theory *13* (1984) 383—399.
[4] Vlachý J.: Czech. J. Phys. *B 35* (1985) 705—708.
[5] Todorov R., Vlachý J.: Czech. J. Phys. *B 36* (1986) 163—166, table 1.
[6] World Development Report 1982. World Bank, Oxford University Press, Oxford, 1982. The Europa Yearbook, Vol. 1. Europa Publ., London, 1984, p. XIV—XVIII.
[7] Vlachý J.: Czech. J. Phys. *B 35* (1985) 801—804.
[8] Stigler S. M.: Ann. Statist. *5* (1977) 1055—1098.
[9] Kovanic P., Novovičová J.: submitted for publication.
[10] Braun T., Glänzel T., Schubert A., Telcs A.: Facts and figures on the publ. output and citation impact of 107 countries as refl. in ISI's database, 1978—1980. Int. Workshop to Assess the Coverage of the Scient. Output of the Third World. Rockefeler Found., Philadelphia, 1985, p. 30.
[11] Schubert A., Glänzel W., Braun T.: Czech. J. Phys. *B 36* (1986) 126—129.
[12] Narin F., Olivastro D.: Czech. J. Phys. *B 36* (1986) 101—106, table 4.
[13] Vlachý J.: Czech. J. Phys. *B 36* (1986).