The use of information theory in atomic spectrochemistry II.* Applicable information contents and information efficiencies

[a]K. ECKSCHLAGER and [b]M. MATHERNY

[a]Institute of Inorganic Chemistry, Czechoslovak Academy of Sciences, CS-250 68 Prague

[b]Department of Chemistry, Faculty of Metallurgy, Technical University, CS-043 85 Košice

Dedicated to Professor Ľ. Treindl, DrSc., in honour of his 60th birthday

The applicable information contents and information efficiency are preferentially given by the relevance of analytical results. However, an unambiguous determination of the coefficients of relevance specific for each element is necessary for numerical calculations. These coefficients may be obtained either on the basis of subjective attribution of "weighing" factor or on the basis of some suitably chosen metrological characteristics such as precision or detectability. Recently some more exact procedures based on "fuzzy" sets have been applied. In this work, all possibilities of relevance applications based on experimental results are examined.

Применимое содержание информаций и их эффективность предпочтительно даны значительностью аналитических результатов. Однозначное определение коэффициентов значительности специфических для каждого элемента нужно для числовых расчетов. Эти коэффициенты могут быть получены или на основании субъективного приписывания «взвешивающего» фактора, или на основании некоторых удобно избранных метрологических характеристик, таких как точность или определимость. Недавно были применены более точные методы основаны на «туманных» множествах. В этой работе рассматривались все возможности значительности применений, основанных на экспериментальных результатах.

The information about chemical composition of a matrix consisting of several elements which can be obtained experimentally, *e.g.* by atomic spectrochemical investigation, has like every information, not only a quantitative but also a subject aspect. As a rule, the use of the information theory in analytical chemistry based on the abstraction from subject aspect and thus only the abundance of final results or the obtained information quantity of a multielemental analysis is to be estimated. However, the relevance of final results is

---

* For Part *I* see Ref. [3].

much more important for a given problem, especially for multielement determination. Up to now, this parameter has been conceived rather qualitatively. Nevertheless, this parameter can be so quantified that it can express the information quantity and, at the same time the contribution of the components important for the solved analytical problem [1]. It has been proved earlier [2—6] that the information contents depend on the properties of the analytical system. The subject feature, on the other hand, is conditioned by the given analytical problem solved by the use of this system. The relevance can be regarded as a contribution of information on a given content of the system.

In this paper it is explained how the useful information contents or information efficiency depend on experimental conditions of the multielement determination and on properties of the atomic spectrochemical process. The quantity named as usable information amount is defined as information contents or information efficiency [3]. The relevance coefficient $k(X_{REL})$ is used as "weight factor". The value of $k(X_{REL})$ is to be considered with regard to the requirements of the analytical problem solved as well as to the importance of the individual analytical elements for the solution of the analytical problem as a whole. The above-mentioned quantities were already applied earlier [3, 4, 7] but, in this study they have been modified for particular valuation of instruments and processes of atomic spectrochemistry. This modification consists in such application to the calculations of usable information contents and information efficiency which is the most appropriate for practical use [7, 8].

Theoretical*

The emission spectrochemical processes enable simultaneous quantitative determination of the number Q of analytical elements with a maximum value $Q_{max} = 10—40$. The analytical signal obtained by an atomic emission spectrochemical process with photographic registration is a two-dimensional information. The stable position of the line signal represents its conditioned part. At the same time, the signal contains the information on the identity of the analytical element (X). The relative intensity of signal $I(X)$ has a random character. It represents the principal value in quantitative analysis since it contains the information on the contents of the analytical element and thus represents the required basic information.

---

*The symbols used in this text are listed on page 633 and are identical with the symbols in paper [3].

In this connection, it has been shown earlier [2, 3] that the a posteriori uncertainty of final results of the multielement determinations and thus the corresponding information contents are conditioned as follows:

a) by the selectivity or by the possibility of resolution of qualitatively different identities,

b) by the relative precision of the concentration determination,

c) as far as the determination of trace elements is concerned, also by the detectability of the given elements.

It depends on the applied methods, on the character of analytical instrumentation as well as analytical procedure, and on analytical problem which of the above-mentioned reasons of uncertainty appear as dominant or whether all three factors participate equally in actual a priori uncertainty. This statement is especially important for the choice of a suitable mathematical model on the basis of which the usable information would be determined.

Usable information measure

At first, it is necessary to explain how the selectivity, the relative precision and the detectability obtained by the atomic emission spectrochemical analysis affect the information quantity. The position of analytical signal, the wavelength of a spectral line is a constant value given by the energy difference between the excited and the ground state of a given analytical element (eqn (*1*))

$$\lambda(X) = hc / \Delta E(X) \tag{1}$$

Despite this, in decoding of spectra it is necessary to regard this value as a random quantity because of errors in measurements and in other experimental parameters. For this reason, we assume that the number of the resolved positions (*m*) in atomic emission spectra is high but finite

$$m = (\lambda(X_{max}) - \lambda(X_{min})) / \Delta\lambda(X) \tag{2}$$

The values of $\lambda(X_{max})$ and $\lambda(X_{min})$ represent for the given experimental arrangement the highest and the lowest registrable wavelengths and the value of $\Delta\lambda(X)$ is primarily related to the reciprocal dispersion of the spectrograph. Therefore this value obeys the inequality $\Delta\lambda(X) \leq 0.01$ nm and the limiting value $m \approx 10^5$ is valid. Every analytical element determined by atomic emission spectrochemical methods usually exhibits more convenient spectral lines, the resolution of spectra is almost always sufficient, and $m \gg Q$. Therefore it is possible to investigate the information contents in such a way that the uncertainty due to insufficient selectivity of the methods [2] may be neglected.

The problem of the uncertainty depending on relative precision of concentra-

tion determination of the main and minor elements as well as on the precision and detectability of the trace elements is much more important [2, 3, 8]. According to established convention the boundary between minor and trace elements is put to the value 10^{-4} %. The parameters of information theory for main and minor elements have been explicitly defined in preceding paper [3]. In case that the value of $s(c_X)$ is not constant for the whole concentration range under consideration, it is necessary to modify the equation for ascertaining the information contents as follows [8]

$$I(p, p_0) = \ln\left(\left(\left(c(X_2) - c(X_G)\right) / s(c_{X, \Sigma})\right) \cdot \left(\left(N^{1/2} / 2t(F, \alpha)\right)\right)\right) \qquad (3)$$

where $s(c_{X, \Sigma})$ stands for the so-called mean standard deviation [4]

$$s(c_{X, \Sigma}) = \left(\left((K - 1)(s(c_X))^2 + (s(c_{X, G}))^2\right) / K\right)^{1/2} \qquad (4)$$

Symbol K denotes the number of calibration concentrations used for construction of the calibration lines.

For some special analytical problems it is imperative to extend the concentration range as far as to the limit of detection $c(X_L)$ (Fig. 1).
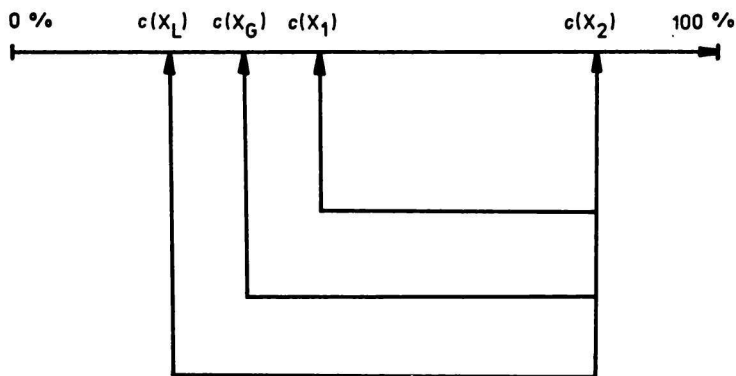

Fig. 1. Illustration of the concentration ranges.

In this case the information contents can be calculated as follows

$$s(c_{X, \Sigma}) = \left(\left((K - 1)(s(c_X))^2 + (s(c_{X, L}))^2\right) / K\right)^{1/2} \qquad (5)$$

According to the Kaiser's principle, $s(c_{X, L, r})$ equals 33.3 % and thus the following standardization holds for the value of $s(c_{X, L})$

$$s(c_{X, L}) = c(X_L) \cdot 0.333 \qquad (6)$$

The value of the measure of information contents can be obtained by summarizing all $(I(p, p_0))_g$ values.

$$\mathrm{MI}(p, p_0) = \sum_{g=1}^{Q} (I(p, p_0))_g \qquad g = 1, \ldots, Q_{\max} \qquad (7)$$

while the individual $(I(p, p_0))_g$ values are to be found by using of eqns (3, 4, and 5). However, it means that $\mathrm{MI}(p, p_0)$ comprises the information contents for all pertinent elements in all occurring concentration ranges, irrespective of their being relevant or irrelevant, for the whole analytical problem.

Relevance of information

Up to now, the relevance of the information obtained experimentally has been only qualitatively estimated. Now we shall try to characterize the relevance quantitatively by using the corresponding function $k(X_{REL})$ of the information on the g-components of unsharp "fuzzy" quantity of the relevant information. The relevance coefficient $k(X_{REL})$ is defined by the unequality $0 < k(X_{REL}) \leq 1$. It is applied as "weight" factor [9] of the information contents that characterizes the relevance

$$\mathrm{MI}^*(p, p_0) = \sum_{g=1}^{Q} (I(p, p_0) \, k(X_{REL}))_g \qquad (8)$$

The relevance coefficient can also be related to the information efficiency [10]

$$\mathrm{ME}^*(p, p_0) = \sum_{g=1}^{Q} (E(X) \, I(p, p_0) \, k(X_{REL}))_g \qquad (9)$$

An actual specific determination of the values of relevance coefficients $k(X_{REL})$ can be based either on partially subjective considerations or on those principles which are frequently used for metrological characteristics of the applied methods and, if needed, on the basic parameters of the social order [11].

A "subjective" solution requires formation of some classes that differ from each other in numerical values of $k(X_{REL})$ and corresponding analytical significance (Table 1). The elements that provide for significant items of information in a given multielement determination with respect to social order must exhibit complete relevance $k(X_{REL}) = 1$. A reduced relevance $k(X_{REL}) = \langle 0.75, 0.50 \rangle$ is to be attributed to the elements with a certain significance or additional information. As to the multielement determinations we must already during development of an analytical method bear in mind that some elements have "unclear" significance or poor information efficiency at the beginning of the method application. That is why these elements are included into the series of analytical elements but they obtain much lower relevance level for final estimation on the basis of information theory ($k(X_{REL}) = 0.25$). The use of relevance

Table 1

Definition of one-dimensional field of relevance coefficients

$k(X_{REL})$	Significance
1.00	Information is absolutely necessary.
0.75	Information is still significant.
0.50	Information significance is either unclear or even contradictory.
0.25	Information may be useful in the future.

coefficients indeed may eliminate the "apparent" high information contents caused by some less important elements. The $k(X_{REL})$ value equal to zero is out of question because the unnecessary elements are not determined.

Another way of establishing the relevance coefficients is based on respecting of some metrological characteristics such as the relative precision of the concentration determination or the detectability [8]. The modified value of relative precision of the concentration determination $s(c_{X,r})$ may be used for the determination of minor elements in which the precision of concentration determination is of especial importance. However this value may be used only in the form without multiplying by 100. For the values $s(c_{X,r}) < 0.10$ the value of relevance coefficients is conventionally regarded as equal to one. As to the values in the range $s(c_{X,r}) \in \langle 1, 0.1 \rangle$ eqn (10) gives the required solution

$$k(X_{REL}) = \log(1 / s(c_{X,r})) \tag{10}$$

The relevance coefficient can also conventionally be considered as equal to one for the analytical methods in which the guarantee limit $c(X_G)$ attains the value of $10^{-4}\%$ at least. On the other hand, the relevance coefficient is always less than one for $c(X_G) > 10^{-4}\%$. For the calculation of a given $k(X_{REL})$ value the following equation is to be recommended

$$k(X_{REL}) = \log((1 / c(X_G)) / 4) \tag{11}$$

The complete possibility is provided by the use of the above-mentioned unsharp "fuzzy" quantities [11, 12] involving the relevance coefficient as a corresponding function, *i.e.* an element specific constant.

The subjective attribution of gradual relevance coefficients (Table 1) can be extended on the basis of the unsharp "fuzzy" quantities by using further classification criteria (Table 2). Other kind of cross-criteria can be defined by quite different meaning. In this case it is possible to use the measure of essentiality or toxicity of trace elements as criteria of valuation. In interpretation of the

632

Table 2

Two-dimensional field of relevance coefficients

Group	Class			
	1.	2.	3.	4.
I.	1.0	0.75	0.50	0.25
II.	0.66	0.50	0.33	0.17
III.	0.33	0.25	0.17	0.08

geological and geochemical data the geochemical "occurrence criteria" or the limitations derived from these values can be employed as cross-criteria.

Symbols

$k(X_{REL})$	relevance coefficient
$I(p, p_0)$	information content
$c(X_1)$, $c(X_2)$	concentration values
$c(X_G)$	limit of guarantee
$c(X_L)$	limit of detection
$s(c_{X, r})$	relative standard deviation of the $c(X_i)$ value
$s(c_{X, \Sigma})$	mean standard deviation
$s(c_{X, L})$	standard deviation of the limit of detection
F	degree of freedom
α	significance level
$MI(p, p_0)$	measure of information content
$ME(p, p_0)$	measure of information efficiency
MI^*, ME^*	the value is corrected for relevance

References

1. Eckschlager, K. and Štěpánek, V., *Analytical Measurement and Information.* Research Studies Press., Letchworth, 1985.
2. Eckschlager, K., *Collect. Czechoslov. Chem. Commun. 47,* 1580 (1982).
3. Matherny, M. and Eckschlager, K., *Chem. Zvesti 38,* 479 (1984).
4. Flórián, K. and Matherny, M., *Fresenius Z. Anal. Chem. 324,* 525 (1986).
5. Danzer, K., Eckschlager, K., and Wienke, D., *Fresenius Z. Anal. Chem. 327,* 312 (1987).

6. Danzer, K., Eckschlager, K., and Matherny, M., *Fresenius Z. Anal. Chem. 334,* 1 (1989).
7. Eckschlager, K. and Matherny, M., *Magy. Kem. Foly. 93,* 564 (1987).
8. Matherny, M., *Wissenschaftliche Beiträge der K. Marx-Universität.* Leipzig, Analytiktreffen 1986. Vol. *II,* p. 490. 1987.
9. Eckschlager, K., *Kem. Kozl. 52,* 141 (1979).
10. Danzer, K. and Eckschlager, K., *Talanta 25,* 725 (1978).
11. Eckschlager, K. and Štěpánek, V., *Chemometrics and Intelligent Laboratory Systems 1,* 273 (1987).
12. Novák, V., *Informačné systémy 6,* 519 (1977).

Translated by M. Matherny