

Literature and Internet Database Mining in a Study About the Word *CHEMOMETRICS*

Rudolf Kiralj*, Márcia M. C. Ferreira rudolf@iqm.unicamp.br

Laboratório de Quimiometria Teórica e Aplicada, Instituto de Química, Universidade Estadual de Campinas, Campinas – SP, 13083-970 BRAZIL

Keywords: bibliometrics and webometrics, linguistics of CHEMOMETRICS, exploratory analysis

The term *chemometrics* was coined by Svante Wold first in Swedish (*kemometri*) in 1971¹ and soon in English for a newly emerging chemical area that employed mathematical and statistical methods in the treatment of chemical data. Since then this term has entered into many languages as a part of common chemical linguistics and scientific publications, and gave name to research groups and societies, meetings and schools, regular faculty courses, companies and software. This work tends to investigate some bibliometric, linguistic and sociological aspects of the term *chemometrics* by means of database mining and chemometric methods.

Three series of database minings were performed. In one, Web of Science was searched for generic form "chemometr*" in title, abstract, or keywords, and in address of publications in the period 1971-2005, for the world and each country. In another database mining, Google and Yahoo search engines were extensively employed to find out all online available languages and related countries that use the word *chemometrics* in national languages as well as in English. The third internet surfing (Google) was directed to determine relative frequencies of the previously found forms for *chemometrics* in English and the national languages.

The word *chemometrics* was found in 48 official languages, in 82 orthographic forms that ranged from only one per language like in Swedish or Portuguese (*quimiometria*) to maximum six in English (*chemometrics*, *chemometry*, *chemiometrics*, *chemiometry*, *chemimetrics*, *chemimetry*). English speaking countries, especially the USA and UK, prefer *chemometrics* among the forms more than other countries. The orthographic forms with pronunciations, notably their *-tri*, *-trics*, *-tria* ending varieties, show interesting geographical patterns in Europe that depend on several items and not only on language groups (Germanic, Romance, Slavic, Baltic etc.).

There were found 76 countries worldwide and 36 in Europe that had participated in 3858 publications with word(s) "chemometr*" in 1975-2005. The number of publications signed by officially called chemometric laboratories, groups and departments from 17 countries is rather modest, being 1189 in 1973-2005. It seems that many researchers do not use the word *chemometrics* in publications and even not in the names of their groups. The geographical distribution of the first data set shows interesting trends in the world and even more in Europe. The distribution curves for No. publications per country (log form) tend to approximate normal distribution curves for the world and every continent in future. These time changes are well observable in Europe, where they are slowed down by political and socio-economic processes in Eastern Europe.

Datasets with five descriptors for European countries and all countries were created: No. publications with "chemometr*" (log form), the first publication date, No. Google hits for "chemometrics" (log form), human development index, and researchers-in-research-and-development index. Principal Component Analysis resulted in two principal components with 86% of the total variance for the two datasets. The countries are distinguished with respect to their chemometric activity and existence of chemometric societies and groups. Hierarchical Cluster Analysis was used also in these analyses. The chemometric activity is not completely described as it does not enumerate chemometric publications without "chemometr*", other chemometric activities, and all inflected forms of *chemometrics* in the national languages. Therefore, the observed trends should be considered as qualitative and general, and country-to-country comparisons should be omitted. It can be said in general that active chemometric life is related to scientific/technological and general progress of a country.

The presented bibliometric and internet-based analyses have shown interesting past and present trends in chemometrics and of *chemometrics* in the world, in the continents and particular countries and languages. These results may be useful for the entire chemometric community.

Acknowledgment: FAPESP

References

¹ Kowalski B.; Brown S.; Vandeginste B. J. *Chemometrics* **1987**, *1*, 1-2.