ACTA SCIENTIARUM POLONORUM *Acta Sci. Pol., Technol. Aliment. 4(1) 2005, 17-25*

ADVANCED STATISTICAL METHODS AS A NEW TOOL FOR DATA ANALYSIS IN FOOD AND NUTRITION SCIENCE*

Małgorzata Darewicz, Jerzy Dziuba
University of Warmia and Mazury in Olsztyn

Abstract. The paper presents a review of literature databases, paying special attention to the use of advanced statistical method as a tool for data analysis in food and nutrition science. Two bibliographic databases, i.e. CAB and FSTA, were searched thoroughly in the study. A dynamic increase in the number of publications based on artificial intelligence studies was observed. A large body of investigations is devoted to the problems of food quality assurance and food authenticity control.

Key words: CAB and FSTA databases, food science, functional properties, multiple regression, neural networks, partial least squares regression, principal component analysis

INTRODUCTION

For many years linear modeling has been universally applied in order to mathematically describe various phenomena and processes. Optimization strategies employed for linear model generation may bring positive results. However, such models do not work in all situations and sometimes lead to wrong conclusions, including the thesis that certain phenomena and processes cannot be described mathematically at all. In such cases advanced statistical methods may be helpful as mathematical models representing nonlinear dependences [Hopke 2003].

Molecular bases of the functional and biological properties of hydrocolloids have been studied at the Department of Food Biochemistry for a long time. Particular attention has been paid to proteins and products of their hydrolysis, i.e. peptides, amino acids and hydrolyzates [Darewicz 2001, Dziuba et al. 2003]. A mathematical description of the protein structure-function interdependence, and determination of its specific nature requires to perform an analysis of two data matrices. One of them should describe mo-

* This work has been financially supported by funds of UWM, project no. 522-0712-0203

Corresponding author – Adres do korespondencji: prof. dr hab. Jerzy Dziuba, Department of Food Biochemistry of University of Warmia and Mazury in Olsztyn, Pl. Cieszyński 1, 10-728 Olsztyn, e-mail: jerzy.dziuba@uwm.edu.pl

lecular properties, and the other – functional properties. Such an analysis can be made using different multivariate statistical methods. These methods can be divided into two categories, i.e. symmetric and asymmetric, depending on the way of treating data matrices [Dijksterhuis 2001]. According to asymmetric methods, data matrices are to be treated differently. These methods consist in building models used for predicting attribute values, e.g. functional properties, on the basis of measurements of molecular properties. Basic asymmetric methods are partial least squares regression and principal component analysis. According to symmetric methods, data matrices are to be treated equally. These methods focus on relationships between data blocks, none of which is the object of prediction. Neural networks may be also applied to modeling the structure-function interdependences, since they are often used for modeling strongly nonlinear data [Wilkinson and Yuksel 1997].

Principal component analysis enables to transform a set of variables describing multivariate observations into a new, reduced set of variables referred to as principal components [Aczel 2000, Morrison 1990]. This transformation goes as follows: the first component explains the largest possible part of data variation, the second components explains the largest possible part of the remaining variation, etc. This procedure is repeated until the desirable percentage of variation is explained. The choice of the number of components is usually arbitrary, but the Keiser's or the Cattell's criteria are applied most often in practice [Rusnak 1999].

Partial least squares regression determines a linear dependence between a set of dependent variables and a set of independent variables through searching for a set of common hidden factors. Partial least squares regression is an expansion of multiple regression. It consists in data transformation into several linear components, which are then used as independent variables [Naes et al. 1996].

Artificial neural networks have been developed as a result of studies in the field of artificial intelligence, especially those relating to models of major brain structures. ANNs use information on the non-linearity of neurons, and the human brain's capabilities, including the ability to learn and memorize. Neural networks are a modeling technique permitting to map very complex nonlinear functions. On the basis of the provided data (dependent and independent variables), they construct regression models or solve problems related to prediction and classification [Tadeusiewicz 2000]. In the case of neural networks, the level of user's theoretical knowledge can be much lower than in that of traditional statistical methods. Neural networks are applied primarily to solving problems connected with classification, control and prediction. Their role in food and nutrition science is constantly increasing due to recent developments in software and computer science.

The paper presents a review of literature databases, paying special attention to the use of advanced statistical method as a tool for data analysis in food and nutrition science. The possibilities offered by principal component analysis, partial least squares regression, multiple regression and neural networks, as well as the limitations of these methods, are also discussed.

MATERIAL AND METHODS

Scientific papers dealing with the application of advanced statistical methods in food and nutrition science were searched for in the following bibliographic databases: FSTA (Food Science and Technology Abstracts), LSC (Life Sciences Collection) and CAB Abstracts (Center of Agriculture and Biosciences International), included in the Ovid Technologies consortium, available on-line at the Main Library of the University of Warmia and Mazury in Olsztyn. The bibliographic base Agricola (National Agricultural Library), available on the Internet, was also used for search. A preliminary analysis of results enabled to select two bibliographic databases, i.e. CAB and FSTA, since they fulfilled the requirements of in-depth presentation of the use of advanced statistical methods in food science.

In order to attain the goal of the study, four main headwords were selected, namely: neural network, principal component analysis, partial least squares regression, multiple regression, and nine additional key words in the field of food science, i.e. food, food science, nutrition, protein, peptide, amino acid, functional properties, foam and emulsion.

Databases were analyzed according to a two-step procedure. At the first stage, each of the four main headwords was searched for individually. At the second stage each of the headwords was connected by the conjunction "and" with each of additional key words. The search was limited to time intervals covered by bibliographic databases.

RESULTS AND DISCUSSION

The databases searched in our study are very popular databases collecting and sharing knowledge about food and natural sciences, used by food chemists, food analysts, food specialists and biologists. The bibliographic database CAB Abstracts, created in 1973, indexes over 11 000 journals on agriculture, forestry and related sciences. The FSTA database includes articles from over 1800 journals on food science and associated disciplines, from 1969 to the present. Another factor increasing the popularity of these databases is that they are widely available via the Internet, in usable forms and free of charge. Sometimes search options are limited, e.g. search results cannot be copied, full texts of articles are unavailable, search is limited to a certain number of publications only, etc.

The database search results obtained on July 1, 2004, are presented in Tables 1, 2, 3 and 4. They show that the number of publications devoted to advanced statistical methods, selected by the query system in the CAB database, was equal to or higher than that of articles in the FSTA database. Authors of the greatest number of publications contained in the CAB Abstracts database applied mainly principal component analysis and multiple regression. The body of literature dealing with neural networks was larger than that of references employing partial least-square regression analysis. This seems interesting, since a dynamic increase in the number of publications based on artificial intelligence studies took place at the beginning of the 1990s, and first reports suggesting the possibility of using artificial neural networks in food-related areas were published in the mid-1990s. The CAB database is a better source of references on advanced statistical methods employed in food and nutrition studies, since their number has always been

greater there than in the FSTA database. This is natural if we compare the coverage of journals abstracted and indexed by both databases. However, the FSTA database provides more information on additional key words, such as functional properties or foam and emulsion, fitting into the scope of scientific interests of the authors of this report, and marking new application ranges of advanced statistical methods. This is probably due to the fact that this database offers insights primarily into food-related disciplines.

Table 1. Number of publications found in literature databases fulfilling the search criteria
Tabela 1. Liczba publikacji spełniających założone kryteria przeszukiwań w literaturowych bazach danych

Literature database Literaturowa baza danych	Number of publications containing the headword ... Liczba publikacji z wyrazem ...									
	headword hasło podstawowe	headword – advanced statistical method and ... hasło podstawowe – zaawansowana metoda statystyczna i ...								
	principal component analysis analiza składowych głównych	food żywność	food science nauka o żywności	protein białko	peptide peptyd	amino acid aminokwas	nutrition żywienie	functional properties właściwości funkcjonalne	foam piana	emulsion emulsja
CAB	4 956	934	177	262	24	59	415	5	3	7
FSTA	1 205	774	174	133	17	54	34	20	13	6

Table 2. Number of publications found in literature databases fulfilling the search criteria
Tabela 2. Liczba publikacji spełniających założone kryteria przeszukiwań w literaturowych bazach danych

Literature database Literaturowa baza danych	Number of publications containing the headword ... Liczba publikacji z wyrazem ...									
	headword hasło podstawowe	headword – advanced statistical method and ... hasło podstawowe – zaawansowana metoda statystyczna i ...								
	partial least square analysis metoda częściowych najmniejszych kwadratów	food żywność	food science nauka o żywności	protein białko	peptide peptyd	amino acid aminokwas	nutrition żywienie	functional properties właściwości funkcjonalne	foam piana	emulsion emulsja
CAB	659	326	100	92	4	23	65	0	0	2
FSTA	676	367	20	81	5	18	5	13	3	3

Table 3. Number of publications found in literature databases fulfilling the search criteria
Tabela 3. Liczba publikacji spełniających założone kryteria przeszukiwań w literaturowych bazach danych

Literature database Literaturowa baza danych	Number of publications containing the headword ... Liczba publikacji z wyrazem ...									
	headword hasło podstawowe	headword – advanced statistical method and ... hasło podstawowe – zaawansowana metoda statystyczna i ...								
	neural network sieć neuronowa	food żywność	food science nauka o żywności	protein białko	peptide peptyd	amino acid aminokwas	nutrition żywienie	functional properties właściwości funkcjonalne	foam piana	emulsion emulsja
CAB	1 860	380	43	40	21	15	102	4	1	1
FSTA	468	275	44	24	0	13	1	1	2	1

Table 4. Number of publications found in literature databases fulfilling the search criteria
Tabela 4. Liczba publikacji spełniających założone kryteria przeszukiwań w literaturowych bazach danych

Literature database Literaturowa baza danych	Number of publications containing the headword ... Liczba publikacji z wyrazem ...									
	headword hasło podstawowe	headword – advanced statistical method and ... hasło podstawowe – zaawansowana metoda statystyczna i ...								
	multiple regression regresja wielokrotna	food żywność	food science nauka o żywności	protein białko	peptide peptyd	amino acid aminokwas	nutrition żywienie	functional properties właściwości funkcjonalne	foam piana	emulsion emulsja
CAB	484	61	10	45	5	3	155	1	1	1
FSTA	33	20	5	4	0	0	9	2	1	1

Most of the papers analyzed in this study are focused on food microbiology, food safety and food quality assurance [Sasic and Ozaki 2001]. For instance, artificial neural networks have been used for the construction of "electronic noses" in order to estimate the microbiological quality of raw materials, food and animal feed [Schnurer et al. 1999]. Another group of problems deals with the possibility to determine the origin of raw materials, to characterize food products and to detect undesirable additives or possible adulteration [Cordella et al. 2002]. Advanced statistical methods have been also applied to identify and detect fraud and adulteration within different cultivars of Spanish bean, previously analyzed by standard electrophoretic techniques [Guerrero et al. 2001]. Due to globalization of European and world trade, the problems of assuring high quality

of foodstuffs are becoming more and more important. The scientific achievements of the employees of the Department of Food Biochemistry, University of Warmia and Mazury in Olsztyn, correspond with this new trend of international research. Our team uses advanced statistical methods for analysis of data acquired by traditional instrumental techniques. At our Department multiple regression equations were used to develop an industrial method for detecting the addition of milk protein preparations to soybean proteins [Dziuba et al. 2004 a, b]. The authors analyzed 57 fractions by RP-HPLC and then used principal component analysis and cluster analysis to select those which statistically significantly ($p < 0.05$) contributed to explaining the variation of the dependent variable in the econometric model tested. The regression equation derived in the study enabled to estimate the level of milk protein preparation in mixtures with soybean protein isolate, but did not permit to determine the type of milk protein preparation added to soybean protein isolate. To achieve this aim, the authors determined a function of third-degree polynomial regression, where independent variable was % area of peaks of all chromatographically separated fractions for particular milk protein preparations added purposefully to a soybean isolate sample. Then the authors selected two unique fractions with specified retention times, the so called "indicators", for each of the four milk protein preparations tested (i.e. calcium caseinate, sodium caseinate, whey protein concentrate and milk protein coprecipitate). The areas of selected peptide peaks ("indicators") were correlated with the percentage of a given milk protein preparation to the highest degree, which allowed to distinguish between a pure soybean protein preparation and a soybean protein preparation containing a milk protein preparation. The results obtained allowed to identify the type of milk protein preparation and its percentage in soybean isolate. In order to reduce (below 5%) the detection limit of milk protein preparation in mixtures with soybean protein isolate, enzymatic hydrolysis by trypsin was performed. Using the results of cluster analysis, a spreadsheet was developed in Microsoft Excel'97 for Windows'98, enabling quantitative detection of milk protein preparations in mixtures with soybean protein preparations even at a level of 1%.

Database search results show that advanced statistical methods, based upon principal component analysis, partial least squares regression or neural networks, can also contribute to characterizing the structure of peptides and food proteins [Lavine and Workman 2002]. For instance, advanced statistical methods have been used to characterize protein hydrolysis products during cheese ripening [Bara-Herczegh et al. 2002].

The problems discussed in the above reference databases concern food and nutrition science, and associated disciplines [Gordon et al. 1998]. Wądołowska et al. [2003] made an attempt to apply advanced statistical methods to analysis of nutritional habits among young people. The use of neural networks as a statistical method in regression model creation enabled these authors to generate a model reflecting real relationships between milk product options and milk consumption frequency. Neural networks were also employed to predict the level of amino acids in food ingredients [Roush and Cravener 1997].

Another crucial research problem is grasping interrelations between molecular and functional properties [Nakai et al. 1986, 1996, Van der Ven et al. 2001, 2002 a]. Nakai et al. [1986, 1996] used regression equations to study the role of protein solubility in a mathematical interpretation of their functional properties. The introduction of surface hydrophobicity and solubility data into a multiple regression model for determining the emulsifying properties of proteins, caused an increase in the determination coefficient

R^2 from 0.78 to 0.92 ($p < 0.001$) [Nakai et al. 1996]. A similar mathematical analysis was also employed by Voutsinas et al. [1983] who studied the emulsifying properties of proteins of various origin following heat treatment. The authors pointed to the possibility of predicting the foam-forming and emulsifying properties, and fat absorption capacity by regression equations taking into account solubility and aliphatic hydrophobicity.

Van der Ven et al. [2001, 2002 a] used partial least squares regression as a statistical tool to develop a model predicting the foam-forming and emulsifying properties of milk protein hydrolyzates on the basis of peptide molecular weights determined by gel chromatography. An alternative to universally applied methods for determining the functional properties of hydrolyzates may be Fourier-transform infrared spectroscopy [Van der Ven et al. 2002 b]. Principal component analysis revealed that whey and casein hydrolyzates prepared with different classes of proteolytic enzymes (i.e. acid, neutral and alkaline) can be effectively distinguished on the basis of FTIR spectra (1800--800 cm^{-1}) [Van der Ven et al. 2002 b]. In addition, Van der Ven et al. [2002 b] applied partial least squares regression to a statistical interpretation of FTIR spectra. A combination of both methods contributed to building a mathematical model to predict bitterness, solubility, emulsifying, and foaming properties of hydrolyzates on the basis of their FTIR spectra. Sörensen and Jepsen [1998] used near-infrared spectroscopy to predict a bitter taste of cheese. A bitter taste of eight samples was predicted using a standard curve model calibrated on the basis of data for 24 samples; the resultant value of the determination coefficient R^2 was 0.28 to 0.59. A bitter taste of cheese was predicted combining data obtained by RP-HPLC and chemical properties (casein content, ratio between unltrafiltration fractions, ratio between hydrophobic and hydrophilic peptides). Regression analysis allowed to explain bitter taste variation in 95% for six samples, and in 59% for 19 samples [Frister et al. 2000].

Arteaga and Nakai [1993] developed artificial neural networks for predicting foam-forming and emulsifying properties, which were correlated with 18 physicochemical parameters of proteins, such as % of α-helix, hydrophobicity, net charge. Taub and Singh [1998] used second- and third-degree Taylor polynomials including variables x, y, their transformations: ln (x), $1/x$, ln (y), $1/y$, and interactions between them. Such equations are commonly applied in practice, e.g. to predict quality deterioration during food storage [Taub and Singh 1998].

CONCLUSIONS

The analysis of literature databases performed in our study indicates a growing interest of researchers in the application of advanced statistical methods in food and nutrition science. It should be emphasized that these methods are widely used in such areas as food quality assurance and control. The range of applications of the most recent of the above techniques, i.e. artificial neural networks, is expanding dynamically. Easy access to modern software and hardware offering high processor capacity facilitates the development of innovative prediction models replacing traditional empirical methods, thus improving food safety and quality.

REFERENCES

Aczel A.D., 2000. Statystyka w zarządzaniu. PWN Warszawa.

Arteaga G.E., Nakai S., 1993. Predicting protein functionality with artificial neural networks: foaming and emulsifying properties. J. Food Sci. 58, 1152-1156.

Bara-Herczegh O., Hprváth-Almássy K., Örsi F., 2002. Application of multivariate methods to identify the indices of secondary proteolysis for Trappist cheese maturity and quality. Eur. Food Res. Technol. 214, 516-520.

Cordella C., Moussa I., Martel A.C., Sbirrazzuoli N., Lizzani-Cuvelier L., 2002. Recent developments in food characterization and adulteration detection: technique-oriented perspectives. J. Agric. Food Chem. 50, 1751-1764.

Darewicz M., 2001. Wpływ enzymatycznej modyfikacji kazeiny-β na jej strukturę i wybrane właściwości funkcjonalne. Rozpr. Monogr. 48. Wyd. UWM Olsztyn.

Dijksterhuis G.B., Piggot J.R., 2001. Dynamic methods of sensory analysis. Trends Food Sci. Technol. 11, 284-290.

Dziuba J., Żbikowska A., Darewicz M., Minkiewicz P., 2003. Molekularne podstawy fizykochemicznych i funkcjonalnych właściwości hydrokoloidów. In: Badania naukowe, przegląd osiągnięć. Red. W. Biesiadka, Wyd. UWM Olsztyn.

Dziuba J., Nałęcz D., Minkiewicz P., 2004 a. Chromatographic identification and determination of commercial milk protein preparation in commercial milk protein preparations in mixtures with soybean protein isolate. Milchwissenschaft 59, 366-369.

Dziuba J., Nałęcz D., Minkiewicz P., Dziuba B., 2004 b. Identification and determination of milk and soybean proteins by chromatography and chemometrical data analysis. Anal. Chim. Acta 521, 17-24.

Frister H., Michaelis M., Schwerdtfeger T., Folkenberg D.M., Sorensen N.K., 2000. Evaluation of bitterness in Cheddar cheese. Milchwissenschaft 55, 691-695.

Gordon S.H., Wheeler B.C., Schudy R.B., Wicklow D.T., Greene R.V., 1998. Neural network pattern recognition of photoacustic FTIR spectra and knowledge-based techniques for detection of mycotoxigenic fungi in food grains. J. Food Prot. 61, 221-226.

Guerrero R.M.R., García R.O., Martínez P.Á., 2001. Dry bean cultivar characterization by isoelectric focusing electrophoresis in polyacrylamide gel. J. Sci. Food Agric. 81, 1126-1131.

Hopke P.K., 2003. The evolution of chemometrics. Anal. Chim. Acta 500, 365-377.

Lavine B.K., Workman J.J., 2002. Chemometrics. Anal. Chem. 74, 2763-2770.

Morrison D.F., 1990. Wielowymiarowa analiza statystyczna. PWN Warszawa.

Naes T., Baardseth P., Helgesen H., Isaksson T., 1996, Multivariate techniques in the analysis of meat quality. Meat Sci. 43: 135-149.

Nakai S., Li-Chan E., Hayakawa S., 1986. Contribution of protein hydrophobicity to its functionality. Nahrung 3-4, 327-336.

Nakai S., Li-Chan E.C.Y., Artega G.E., 1996. Measurement of surface hydrophobicity. In: Methods of testing protein functionality. Red. G. M. Hall. Chapmann London.

Roush W.B., Cravener T.L., 1997. Artificial neural network prediction of amino acid levels in feed ingredients. Poult. Sci. 76, 721.

Rusnak Z., 1999. Metoda składowych głównych. In: Statystyczne metody analizy danych. Wyd. AE Wrocław.

Sasic S., Ozaki Y., 2001. Short-Wave Near-Infrared spectroscopy of biological fluids. 1. Quantitative analysis of fat, protein and lactose in raw milk by partial least-square regression and band assignment. Anal. Chem. 73, 64-71.

Schnurer J., Olsson J., Borjesson T., 1999. Fungal volatiles as indicators of food and feeds spoilage. Fungal. Genet. Biol. 209, 17-23.

Sörensen L.K., Jepsen R., 1998. Assessment of sensory properties of cheese by near-infrared spectroscopy. Int. Dairy J. 8, 863-871.

Tadeusiewicz R., 2000. Drogi i bezdroża w badaniach naukowych. In: Statystyka w badaniach naukowych. Polska wersja STATISTICA Neural Networks. Seminaria, Warszawa.
Taub A.I., Singh R.P., 1998. Food storage stability. CRC Press LLC Boca Raton USA.
Wądołowska L., Cichon R., Słowińska M.A., 2003. The implementation of advanced exploration techniques in youth nutritional status evaluation. Pol. J. Food Nutr. Sci. 53, 63-68.
Van der Ven C., Gruppen H., de Bont D.B.A., Voragen A.G.J., 2001. Emulsion properties of casein and whey protein hydrolysates and the relation with other hydrolysate characteristics. J. Agric. Food Chem. 49, 5005-5012.
Van der Ven C., Gruppen H., de Bont D.B.A., Voragen A.G.J., 2002 a. Correlations between biochemical characteristics and foam-forming and –stabilizing ability of whey and casein hydrolysates. J. Agric. Food Chem. 50, 2938-2946.
Van der Ven C., Muresan S., Gruppen H., de Bont D.B.A., Merck K.B., Voragen A.G.J., 2002 b. FTIR spectra of whey and casein hydrolysates in relation to their functional properties. J. Agric. Food Chem. 50, 6943-6950.
Voutsinas L.P., Cheung E., Nakai S., 1983. Relationships of hydrophobicity to emulsifying properties of heat denatured proteins. J. Food Sci. 48, 26-32.
Wilkinson C., Yuksel D., 1997. Using artificial neural networks to develop prediction models for sensory-instrumental relationships, an overview. Food Quality Preface 8, 439-445.

ZAAWANSOWANE METODY STATYSTYCZNE JAKO NOWE NARZĘDZIA WYKORZYSTYWANE W ANALIZIE DANYCH W NAUCE O ŻYWNOŚCI I ŻYWIENIU

Streszczenie. W pracy przedstawiono wyniki analizy literaturowych baz danych w aspekcie możliwości zastosowań zaawansowanych metod statystycznych w analizie danych w nauce o żywności i żywieniu. Ostatecznie do przeszukiwań wytypowano abstraktowe bazy danych CAB i FSTA. Zaobserwowano dynamiczny wzrost liczby prac wykorzystujących sztuczne sieci neuronowe w badaniach poświęconych żywności i żywieniu. Zwraca uwagę duża liczba publikacji poświęconych zagadnieniom związanym z zapewnieniem jakości zdrowotnej żywności oraz jej autentyczności.

Słowa kluczowe: analiza składowych głównych, bazy CAB i FSTA, nauka o żywności, regresja metodą cząstkowych najmniejszych kwadratów, regresja wielokrotna, sieci neuronowe, właściwości funkcjonalne

Accepted for print – Zaakceptowano do druku: 21.03.2005 r.

***For citation – Do cytowania**: Darewicz M., Dziuba J., 2005. Advanced statistical methods as a new tool for data analysis in food and nutrition science. Acta Sci. Pol., Technol. Aliment. 4(1), 17-25.*