Published in International Dairy Journal 15, issues 6-9, 557-562, 2005
which should be used for any reference to this work

Geographic origin of European Emmental. Use of discriminant analysis and artificial neural network for classification purposes

L. Pillonel[a], U. Bütikofer[a], H. Schlichtherle-Cerny[a,*], R. Tabacchi[b], J.O. Bosset[a]

[a]*Agroscope Liebefeld-Posieux (ALP), CH-3003 Bern, Switzerland*
[b]*University of Neuchâtel, CH-2007 Neuchâtel, Switzerland*

Abstract

The goal of this work was to classify European Emmental cheeses according to their geographic origin using analytical approaches. Twenty-five analytical parameters (factors) were measured in 183 samples. Results were combined by multivariate statistical analysis. Discriminant analysis (DA) and an artificial neural network (ANN) delivered similar results when all regions and factors were included; 95% and 91%, respectively, of the samples were correctly classified in the validation procedure. To reduce the analytical costs and the risk of overfitting, a DA based on a selection of only 11 factors was calculated. In this case, the Jackknifed validation delivered 95% correct assignments. Finally, a system was optimised to discriminate between the Swiss samples and cheeses from other regions. Building a new model for each of the six pairs, Switzerland vs. another region, 100% correct classification could be achieved for the Swiss samples.

Keywords: Artificial neural network; Discriminant analysis; Emmental cheese; Authenticity; Chemometric data analysis

1. Introduction

Fraud detection in foods often requires a highly accurate characterisation of the product including the use of many different analytical tools. Interesting reviews on food authenticity have been published either oriented towards techniques (Cordella, Moussa, Martel, Sbirrazzuoli, & Lizzani-Cuvelier, 2002; Gremaud, Karlen, & Hulliger, 2002) or food matrices (Dennis, 1998). The determination of the geographic origin of a foodstuff is a difficult task, especially in foods such as cheese that are biochemically and microbiologically dynamic and which undergo changes during ripening. As a consequence, the data from the selected analytical techniques must often be combined by multivariate analysis, also known as chemometrics.

Chemometrics can be defined as the application of mathematical and statistical methods to maximise the chemical information extracted from data. Chemometrics are powerful tools finding applications in various domains covered by published reviews (e.g., Lavine, 1998, 2000; Lavine & Workman, 2002). Pattern recognition is a specific application of chemometrics which occupies the attention of chemists involved in the fight against food fraud. Two comprehensive review articles focusing on chemometrics for authentication and classification of food products were published recently (Tzouros & Arvanitoyannis, 2001; Arvanitoyannis & van Houwelingen-Koukaliaroglou, 2003). Techniques such as principal component analysis (PCA), discriminant analysis (DA), principal component regression (PCR), partial least square (PLS), artificial neural network (ANN) are commonly used for authentication purposes.

*Corresponding author. Tel.: +41 31 323 8167; fax: +41 31 323 8227.
E-mail address: hedwig.schlichtherle-cerny@alp.admin.ch (H. Schlichtherle-Cerny).

A crucial point of pattern recognition is the validation of the model. In an overfitted model, classification into categories may superficially appear satisfactory, but is in fact not statistically significant. The following simple rule-of-thumb should therefore be applied before each classification: the number of factors (variates) included in the model should not exceed $(n-g)/3$, where n is the number of observations and g the number of categories fixed (Defernez & Kemsley, 1997). Moreover, no classification should be carried out without cross-validation. A first possible procedure of cross-validation is to assign each observation at random to either a *training* or a *validation* set. The training set is only used to obtain a model, which is then applied in a second step to the validation set. In DA, typical training and validation sets may contain 66% and 33% of the available observations, respectively. In ANN, training sets may contain up to 80% of cases. If the results between training and validation differ strongly, the model is overfitted or insufficiently adapted. An alternative procedure is the *leave-one-out* or *Jackknifed* validation. This is performed by omitting one observation at a time from the data set and using the remaining data set to obtain a model, which is then applied to the omitted observation. This is repeated n times, excluding each observation in turn and reintroducing the previously omitted observation. The results for the excluded observations are only then assessed. Once again, if normal set and Jackknifed validation diverge, the model is overfitted. Both validation procedures seemed to deliver comparable results (Defernez & Kemsley, 1997).

The present paper deals with the use of chemometrics to determine the geographic origin of Emmental cheese. Pattern recognition methods such as DA and ANN were applied to check a possible classification by region of the cheese samples. During a 3-yr project, promising analytical techniques were selected and applied to 183 Emmental cheese samples from Europe. A more detailed description of the project as well as the corresponding individual results and univariate statistics are presented elsewhere (Pillonel et al., 2005).

2. Materials and methods

2.1. Origin of samples and analytical methods

Sample selection, treatment and characteristics as well as analytical methods used are described by Pillonel et al. (2005). In short, the 183 samples originated from the following 7 regions of Europe: Western Austria (A), Switzerland (CH), South Germany (D) and France Savoie (FR) made using raw milk, and Finland (FI), France Brittany (FT_b) and France East-Central (FT_e) made using thermised milk. The 25 investigated factors finally retained as significant were the following: volatile organic acids formic (C1), acetic (C2), propionic (C3), *n*-butyric (C4) and *n*-caproic (C6), total nitrogen (TN), water-soluble nitrogen (WSN) and 12%-TCA soluble nitrogen (TCA-SN), sodium chloride, D- and L-lactate, succinate and pyruvate, L-leucine-aminopeptidase (LAP) activity, pH, enterococci (ECOC) and obligately heterofermentative lactobacilli (OHL), sodium, copper, magnesium, zinc as well as the stable isotope ratios δ^2H, δ^{13}C, δ^{15}N and δ^{34}S. *Lb. helveticus* data were not retained for modelling because the data were not quantitated. However the presence/absence analysis carried out still delivered most interesting results (Pillonel et al., 2005).

The 20 samples analysed in the preliminary study (Pillonel et al., 2005) were also integrated into the current work. LAP activity and $^{34}S/^{42}S$ isotope ratio had however not yet been investigated in the preliminary study. To allow the software to work correctly, missing values were replaced by the average values of the corresponding category.

2.2. Discriminant analysis

DA on the correlation matrix was performed using Systat for Windows version 9.0 (SPSS Inc., Chicago, USA). An automatically stepwise backward elimination was carried out to select the best variates. In this way, only the most significant factors out of the 25 were used in the corresponding models. Results were validated using the leave-one-out cross-validation (Jackknifed classification matrix).

2.3. Artificial neural network

A feed forward neural network trained by back-propagation was calculated using the software S-Plus (Insightful, Seattle, USA). Briefly, "neurons" were sorted into three layers: input, hidden and output layer (Fig. 1). There were as many input "neurons" as factors

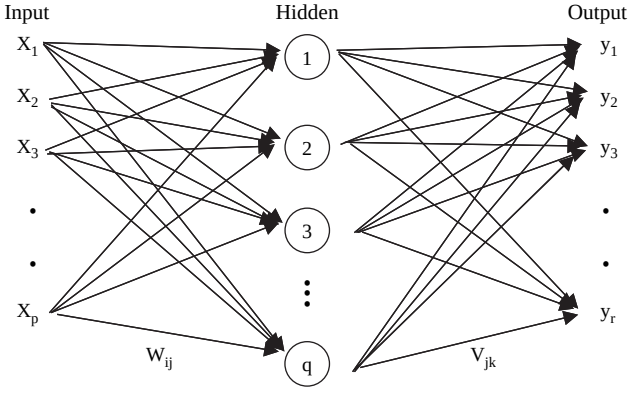

Fig. 1. Schematic representation of the structure of an ANN.

and as many output "neurons" as categories. The input variables were standardised. "Neurons" were connected to all the previous layers by weighted connections. In each "neuron", the sum of the weighted signals was calculated and when it exceeded a certain threshold, it was processed by a so-called transfer function

$$a_k = v_{0k} + \sum_{j=1}^{q} v_{jk}\phi\left[\sum_{i=1}^{p} w_{ij}x_i + w_{0j}\right],$$

(where ϕ is the logistic function) and sent to all "neurons" in the next layer. Direct relation between input and output layer were not allowed. A training dataset containing 66% of the observations ($n = 127$) was randomly selected using a binomial distribution. The remaining data ($n = 56$) were used for validation. In the training sequence, the output of the network was compared to known values and errors were back-propagated to the hidden and input layers to adjust the weights and minimise the error step by step using the method of the steepest descent. The procedure was repeated until the errors between the output and known values were minimised. Several parameters had to be fixed at the beginning of the procedure, e.g., number of hidden "neurons", start weights w_{ij} and v_{jk} (Fig. 1) and weight decay λ. Their selection was optimised as described in the results.

3. Results

Unsupervised classification techniques such as PCA are valuable tools for detecting natural grouping in a set of data. These statistical instruments were used in a preliminary study including only 20 samples to help selecting the analytical methods with the best discriminating potential (Pillonel, Tabacchi, & Bosset, 2003). As here, the group assignment for the collected data is known, and the data set is big enough, only trained (or supervised) classification techniques were considered. DA is a simple and well-understood method of achieving a group assignment (Kaufmann, 1997). The classification was based on Mahalanobis distances and a confidence level is available for each observation. Three approaches with different group assignments were first compared using DA. In the first approach, all regions were included in the model in a single step. In the second approach, only two categories were considered; i.e., a selected region versus the remaining regions pooled in a second category. In the last approach, one region was selected and compared with the other regions considered one by one.

Among the numerous statistical methods available for pattern recognition, ANN was also tested for classifying the samples in a single step. ANN has proven to be a powerful tool for large data sets. The most popular ANN configuration is the back-propagation network (BPN). BPN may give better results than PLS or PCR (Horimoto, Lee, & Nakai, 1997).

3.1. Model 1: all regions considered simultaneously using DA

An excellent classification leading to only two CH samples misclassified was obtained by including 18 factors in a DA. In the Jackknifed classification, nine samples were misclassified, indicating a certain degree of overfitting. Moreover, the analysis of so many factors would cost approximately € 850 per sample. However, reducing the model to 11 factors (C3, pH, WSN, D-lactate, succinate, LAP, Cu, δ^2H, δ^{13}C, δ^{15}N and δ^{34}S), the costs could be reduced to € 650, with 96% (95% in the Jackknifed validation) overall correct assignment still being achieved. One Austrian, five Swiss, two German and one French (FR) samples were misclassified in the Jackknifed validation (Table 1). The probabilities for Swiss membership of the misclassified Swiss samples were all under 0.23. Values equal to or greater than 0.50 ensure a classification as Swiss samples. This highlights the fact that the five misclassified samples showed unusual properties for Swiss Emmental. A pertinent visualisation of the results is difficult due to their multidimensional character. Fig. 2 shows the canonical scores of the first four dimensions. The discrimination is then apparent.

By a further reduction to the factors D-lactate, δ^{13}C, δ^{15}N and δ^{34}S, only the regions distant from one another were correctly separated (FT_b, FT_e, FI). This illustrates the effectiveness of stable isotope ratios for differentiating complex food products that originate from distant geographic origins.

3.2. Model 2: all regions considered simultaneously using ANN

A computer-generated backward elimination of less significant factors as done for DA is not possible here.

Table 1
Jackknifed classification matrix of all observation using 11 factors[a]

	A[b]	CH	D	FI	FR	FT_b	FT_e	%Correct
A	14	0	1	0	0	0	0	93
CH	3	65	2	0	0	0	0	93
D	0	0	21	0	1	0	1	91
FI	0	0	0	12	0	0	0	100
FR	1	0	0	0	30	0	0	97
FT_b	0	0	0	0	0	19	0	100
FT_e	0	0	0	0	0	0	12	100
Total	18	65	24	12	31	19	13	95

[a]Propionate, pH, water-soluble N, D-lactate, succinate, LAP, Cu, δ^2H, δ^{13}C, δ^{15}N and δ^{34}S.

[b]For explanation of cheese codes, see text.

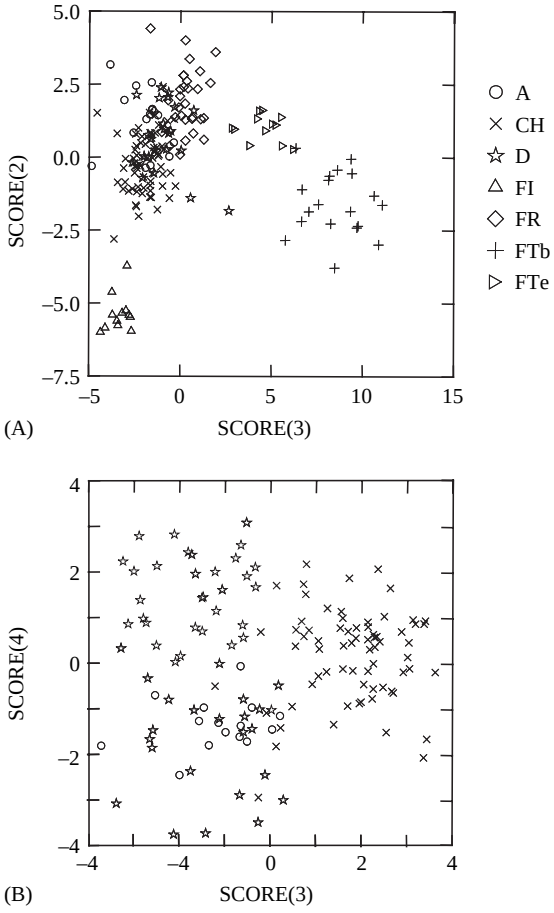

Fig. 2. Canonical scores of the DA using 11 factors: (A) discrimination between the categories FI, FT_b and FT_e using the 1st and 2nd dimensions; (B) partial discrimination between A, CH, D, and FR in the 3rd and 4th dimensions. Only these four categories are represented. For cheese codes, see text.

Table 2
Validation matrix ($\frac{1}{3}$ observations) using all 25 factors in an ANN

	A[a]	CH	D	FI	FR	FT_b	FT_e	%Correct
A	3	0	2	0	0	0	0	60
CH	0	24	0	0	0	0	0	100
D	0	0	4	0	0	0	0	100
FI	0	0	0	6	0	0	0	100
FR	0	0	1	0	11	0	0	92
FT_b	0	0	0	0	0	1	2	33
FT_e	0	0	0	0	0	0	2	100
Total	3	24	7	6	11	1	4	91

[a]For explanation of cheese codes, see text.

Table 3
Validation matrix ($\frac{1}{3}$ observations) using all 25 factors in discriminate analysis

	A[a]	CH	D	FI	FR	FT_b	FT_e	%Correct
A	4	0	1	0	0	0	0	80
CH	0	23	0	0	1	0	0	96
D	1	0	3	0	0	0	0	75
FI	0	0	0	6	0	0	0	100
FR	0	0	0	0	12	0	0	100
FT_b	0	0	0	0	0	3	0	100
FT_e	0	0	0	0	0	0	2	100
Total	5	23	4	6	13	3	2	95

[a]For explanation of cheese codes, see text.

Therefore, all 25 factors were used in the model (input "neurons"). One difficulty of ANN is the arbitrary choice of the initial weights, the weight loss and the number of "neurons" in the hidden layer. As the number of observations was modest, the number of hidden "neurons" was kept relatively low. Various models with a number of "neurons" between 9 and 20, initial weights in the ranges [−0.5; 0.5], [−0.7; 0.7] and [−1; 1] as well as a weight loss of 0.01 or 0.001 were compared. Out of the 72 models tested, the one with 19 hidden "neurons", weight loss of 0.01 and initial weights in the range [−0.5; 0.5] gave the best results. All samples were correctly classified in the training set. In the validation set, five samples were misclassified (Table 2). Two A samples and one FR sample were put in the D group, two FT_b samples were put in the FT_e group.

The distribution of the training/validation sets was sometimes far from 2:1 for the categories with fewer observations. This explains the very low classification rate in regions A and FT_b, though only two samples from each group were misclassified (Table 2). CH was the group with the maximum number of samples and, therefore, the most dependable for use in an ANN. For the other categories, the number of samples was somewhat too small. All CH samples were correctly classified so that the current model can be said to be quite reliable for determining if an unknown sample originates from Switzerland or not. The actual major drawback for such a discrimination is the cost of measuring the 25 factors.

To get a direct comparison with model 1, a DA was carried out with the identical training/validation set as used in ANN. In the training model, only one CH sample was misclassified as D. In the validation set, one A, one CH and one D sample were wrongly classified (Table 3). The results of both ANN and DA were therefore comparable. The better classification obtained for CH with ANN suggested that the performances of the technique could be enhanced with a larger database.

3.3. Model 3: one selected region vs. the others pooled using DA

Each region taken individually one after the other was compared to all others pooled into one category. For

certain regions, this may allow a better discrimination and/or lower costs to be achieved. In this and the following model, the Jackknifed classification matrix was used for validation.

For FI cheeses, the separation was trivial. The factors C3, LAP, L-lactate, pyruvate, Zn, and $\delta^{13}C$ made possible a perfect separation even in the Jackknifed validation. For FR cheeses, 97% correct classification was achieved using the factors C1, C2, TN, Zn, $\delta^{15}N$, $\delta^{34}S$. The FR samples were correctly assigned whereas two samples from each of the categories CH, D and FT_e were assigned to the FR category. Only a small improvement was achieved for A cheeses. Using 13 factors (C2, C4, C6, NaCl, pH, TCA-SN, WSN, OHL, pyruvate, Cu, Zn, $\delta^{2}H$, $\delta^{34}S$), all Austrian Emmentals were correctly classified and one sample from each of the categories CH, FT_e and FR were misclassified. There was a lack of homogeneity amongst the samples from Austria. This is partly due to the diversity of their origin (Vorarlberg, Salzburg) and manufacture (copper vs. stainless-steel vats). A small improvement was also achieved for CH samples. Using the 12 factors C3, NaCl, pH, TCA-SN, WSN, enterococci, OHL, LAP, succinate, Mg, $\delta^{15}N$ and $\delta^{34}S$, only three CH samples were misclassified and no sample was wrongly classified as Swiss.

For the remaining regions FT_e, FT_b and D, the two-category approach did not improve the discrimination, regardless of the number of factors required, or the percentage of correct classification.

3.4. Model 4: Switzerland vs. the others taken one by one using DA

In this approach, a given region (e.g., Switzerland) is compared and confronted to the others considered, one-by-one. An independent model using a specific set of factors is therefore created for each pair. This approach was applied as an example to the Switzerland region to determine if any improvement could be achieved in comparison with both preceeding models. A stepwise backward elimination was first carried out for each pair. Then the factors were compared and manually adapted in order to minimise the number of factors needed for all pairs. At the end, 15 analytical variates (C2, C3, pH, TN, OHL, ECOC, D- and L-lactate, succinate, pyruvate, LAP, $\delta^{2}H$, $\delta^{13}C$, $\delta^{15}N$ and $\delta^{34}S$) were retained. The factors required for each pair are listed in Table 4. A 100% correct classification was achieved in all pairs and also in the Jackknifed validation. For this, the origin of the non-Swiss samples must, of course, be known to apply the correct model. For a sample whose origin is absolutely unknown, each of the six models have to be run. If the sample is always included in the group "Switzerland", it is a Swiss sample. If not, it is non-Swiss, but it is not possible to precisely determine its origin. The one-by-one approach is therefore complementary to the global approach. The former makes it possible to check, with a high confidence level, if the sample is of Swiss origin or not and the latter gives reliable general information about geographic origin.

Table 4
Factors used in model 4

CH vs.	A[a]	D	FI	FR	FT_b	FT_e
C2				X		X
C3	X			X		X
pH		X				
TN		X				
OHL	X			X		
ECOC		X		X		
D-lactate	X	X	X			
L-lactate	X					
Succinate	X	X				
Pyruvate		X	X			
LAP	X	X	X			
$\delta^{2}H$	X					
$\delta^{13}C$			X		X	X
$\delta^{15}N$	X			X	X	X
$\delta^{34}S$	X		X		X	

[a]For explanation of cheese codes, see text.

4. Conclusion

An attempt was made to classify 183 Emmental cheese samples selected from seven European regions according to their seven geographic origins. A maximum of 25 factors (analytical parameters) was available for multivariate analyses. DA and ANN delivered comparable results when all factors were used. In the training set, 99–100% correct classification was achieved, whereas in the validation set, rates between 91% and 95% were found. The size of the database was, however, somewhat too small for ANN to develop its whole power. A further drawback of ANN is its "black box" character. It is not possible to interpret any result or find any relationship between input and output. Hence, it is difficult to reduce the number of factors, costs and the risk of overfitting by selecting the most appropriate factors.

The latter operation is easily carried out in DA using stepwise backward elimination. A new model was optimised using only 11 factors (C3, pH, WSN, D-lactate, succinate, LAP, Cu, $\delta^{2}H$, $\delta^{13}C$, $\delta^{15}N$ and $\delta^{34}S$) leading to 95% correct classifications in the Jackknifed validations. In a two-group approach, samples of one category were compared with all others pooled in a second category. Only slight improvements were achieved this way. A further two-group approach was tested for optimising the separation of the Swiss Emmental only. For each of the six pairs, Switzerland vs. another region, a new model was built. In this way, it

was possible to achieve 100% correct identification for the Swiss samples in the Jackknifed validation using 15 factors (C2, C3, pH, TN, OHL, ECOC, D- and L-lactate, succinate, pyruvate, LAP, δ^2H, δ^{13}C, δ^{15}N and δ^{34}S).

The analytical parameters selected over the 3-yr project, combined with DA were therefore able to assign unknown Emmental samples to their geographic origin with high confidence.

Acknowledgements

The authors thank the Swiss Federal Office of Public Health for its financial and logistical support, G. Urbach and G. Gremaud for their reviewing of the publication as well as J.F. Chamba (ITFF, France), M. Carl (MUVA, Germany), H. Spring (Emmentaler Switzerland™) and H. Jatila (Valio Ltd., Finland) for collecting the samples.

References

Arvanitoyannis, I. S., & van Houwelingen-Koukaliaroglou, M. (2003). Implementation of chemometrics for quality control and authentication of meat and meat products. *Critical Reviews in Food Sciences and Nutrition*, *43*, 173–218.

Cordella, C., Moussa, I., Martel, A.-C., Sbirrazzuoli, N., & Lizzani-Cuvelier, L. (2002). Recent developments in food characterisation and adulteration detection: Technique-oriented perspectives. *Journal of Agricultural and Food Chemistry*, *50*, 1751–1764.

Defernez, M., & Kemsley, E. K. (1997). The use and misuse of chemometrics for treating classification problems. *Trends in Analytical Chemistry*, *16*, 216–221.

Dennis, M. J. (1998). Recent developments in food authentication. *Analyst*, *123*, 151R–156R.

Gremaud, G., Karlen, S., & Hulliger, K. (2002). Analytical methods for the authentication of meat and meat products: Recent developments. *Mitteilungen aus dem Gebiete der Lebensmitteluntersuchung und Hygiene*, *93*, 481–501.

Horimoto, Y., Lee, K., & Nakai, S. (1997). Classification of microbial defects in milk using a dynamic headspace gas chromatograph and computer-aided data processing. 2. Artificial neural network, partial least-squares regression analysis, and principal component regression analysis. *Journal of Agricultural and Food Chemistry*, *45*, 743–747.

Kaufmann, A. (1997). Multivariate statistics as a classification tool in the food laboratory. *AOAC International*, *80*, 665–675.

Lavine, B. K. (1998). Chemometrics. *Analytical Chemistry*, *70*, 209R–228R.

Lavine, B. K. (2000). Chemometrics. *Analytical Chemistry*, *72*, 91R–97R.

Lavine, B. K., & Workman, J. (2002). Chemometrics. *Analytical Chemistry*, *74*, 2763–2769.

Pillonel, L., Badertscher, R., Casey, M., Meyer, J., Rossmann, A., Schlichtherle-Cerny, H., Tabacchi, R., & Bosset, J. O. (2005). Geographic origin of European Emmental cheese: Characterisation and descriptive statistics. *International Dairy Journal*, in press, doi:10.1016/j.idairyj.2004.07.028.

Pillonel, L., Tabacchi, R., & Bosset, J. O. (2003). Analytical methods for the determination of the geographic origin of Emmental cheese. Summary of a screening study. *Mitteilungen aus dem Gebiete der Lebensmitteluntersuchung und Hygiene*, *94*, 60–69.

Tzouros, N. E., & Arvanitoyannis, I. S. (2001). Agricultural produces: Synopsis of employed quality control methods for the authentication of foods and application of chemometrics for the classification of foods according to their variety or geographical origin. *Critical Reviews in Food Sciences and Nutrition*, *41*, 287–319.