Determination of the Geographical Origin of Wines by Methods of Multivariate Data Analysis

U. Römisch1, D. Vandev2, A. Klimmek3 and R. Wittkowski3

1 PROBLEM:
The responsible wine controlling authorities often are confronted with products which are not correctly marked. To find out such adulterations of wines the identification of their geographical origin is of great interest to wine consumers and producers.

Two questions are to be answered:
• Which chemico-analytical variables are important for the determination of the geographical origin (country) of wines and are there differences between red and white wines?
• How many wines can be correctly classified into the groups of countries? Will be new wines correctly classified on the base of the obtained model?

2 DATA:
The study was carried out by 144 wine samples (red and white wines) from 5 East European countries:
• Bulgaria: 38
• Romania: 28
• Hungary: 39
• Macedonia: 29 and
• Moldova: 10.

The 79 measured variables could be classified into 6 chemical groups:
- Biogenic amines (BA): 9
- Inorganic parameters (IP): 6
- Rare earth elements (REE): 15
- Isotope ratios (IR): 7
- General parameters (GP): 33 and

3 STEPS of DATA ANALYSIS:
- Data Management
  - Structure of data, data control, missing values, scaling, transformations, outlier identification
- Distribution of the variables
  - Box-Whisker plots, histograms, distribution fitting
- Variance Analysis (Comparison of Countries)
  - Multiple Box-Whisker plots, ANOVA, multiple comparisons, MANOVA
- Correlation Analysis
  - Correlations between all variables
  - Correlations between groups of variables
  - Factor Analysis
  - Extraction and interpretation of latent factors
  - Extraction of factors restricted to chemical groups

⇒ Reduction of the data matrix

4 RESULTS:
• 96% of the wines could be classified correctly by a discriminant model with 17 parameters. The models for red and white wines were different.
• The obtained model was tested by 10 new wines. 80% of the new wines could be classified correctly.
• The total data set of 144 wine samples was used to fit a new discriminant model with 19 (17+2) parameters.

DATA MANAGEMENT:
1. The data were normally or lognormally distributed.
2. The following variables were to be deleted from the data matrix:
   • Variables which were calculated from others (6 v.)
   • Variables without significant mean differences between the countries (21 v.)
   • Variables which were highly correlated with others (11 v.)
   • Variables which were considered as not so important by the wine specialists (13 v.).
3. Wines which were outside of quality standards were to be deleted from the data matrix (10 wines).

⇒ Reduction of the data matrix to 33 variables and 134 wine samples.