Q01124 Final report: Assessing the origin of wine using existing compositional information.

FINAL REPORT-DRAFT

Report Number	001
Project Manager	Dr. Adrian Charlton
Authors	Helen Grundy Dr. Adrian Charlton
Date	23/02/2009
Customer	Food Standards Agency, UK
Customer's Project No	Q01124
FERA Project No	R6NG
Principal Workers	Paul Brereton, Adrian Charlton, Michal Daszykowski, Helen Grundy, Ivana Stanimirova, Beata Walczak, Oliver Wardman,

fera

The Food and Environment Research Agency

Sand Hutton
York, Y041 1LZ

Tel: 01904 462513
Fax: 01904 462133
Email: adrian.charlton@fera.gsi.gov.uk

1. EXECUTIVE SUMMARY

The UK currently lacks a system for the evaluation of wine origin by analytical methods. The aim of this project is to construct classification models that are capable of predicting the origin of wine, from minimum analytical information, defining the vintage and grape variety of a wine of a given origin with high sensitivity and satisfactory specificity. The relevant analytical data includes eighty three parameters including the levels of components such as trace and rare earth elements, organic and aromatic compounds, alcohol content and glycerol levels in addition to isotope ratios. The wine data used was previously generated during the EU/FSA Third Countries wine project (Q01063, G6RD-CT-2001-00646-WINE-DB).

The wines used in this study were from Romania, Hungary, the Czech Republic and South Africa and nine grape varieties were included. The main objectives of the project were to build models to discriminate the wines according to vintage and also according to grape variety. In order to achieve this, all relevant data was collated in a searchable portal which was built as part of the project. A range of modelling methods, including linear and non-linear methods, were compared for sensitivity and specificity. A final objective was to report on how wine testing is performed elsewhere in Europe.

Wines of a given origin could be distinguished according to grape variety with a high degree of confidence. Wines could also be characterised according to vintage with a high degree of confidence. A combination of up to forty three variables is used to discriminate any one grape from the other grapes in this survey according to vintage. Thirty six variables are required to differentiate grapes according to variety.

Results obtained in this study are strong evidence that, using the chemical data stored in the wine data bank, discrimination of vintages and some grape varieties is possible, when multivariate chemometric methods using selected explanatory variables are applied.

2. CONTENTS

3. GLOSSARY

BfR	Bundesinstitut für Risikobewertung (Federal Institute for Risk Assessment, Germany)
$^{13}C/^{12}C$	The ratio of the carbon isotope with atomic mass 13 to the carbon isotope with atomic mass 12
$\delta^{13}C$	$(R_{sample} - R_{reference\ standard} / R_{standard}) \times 1000$, where $R = {}^{13}C/{}^{12}C$
CycDs	Cyclic diglycerols
D/H	The ratio of deuterium,the hydrogen isotope with atomic mass of 2, to the hydrogen isotope with atomic mass of 1
Ethanol $(D/H)_I$	The ratio of deuterium to hydrogen, when the deuterium ion is positioned on Carbon 1 in the ethanol molecule
Ethanol $(D/H)_{II}$	The ratio of deuterium to hydrogen, when the deuterium ion is positioned on Carbon 2 in the ethanol molecule
Fera	Defra Food and Environment Research Agency
L	litre
mg	milligram
3-MPD	3-methoxy-1,2-propanediol
ng	nanogram
$^{18}O/^{16}O$	The ratio of the oxygen isotope with atomic mass of 18 to the carbon isotope with atomic mass of 16
$\delta^{18}O$	$(R_{sample} - R_{reference\ standard} / R_{standard}) \times 1000$, where $R = {}^{18}C/{}^{16}C$
OIV	Organisation Internationale de la Vigne et du Vin (International Organisation of Vine and Wine)
Na_excess	Excess sodium, compared to chlorine, calculated as: 0.65*Cl - Na
SQL	Structured Query Language

Modelling methods:

CART	Classification and regression trees
DPLS	Discriminant partial least squares
K-DPLS	Kernel discriminant partial least squares
NFS	Neuro-fuzzy systems
PCA	Principal Component Analysis
R-PCA	Robust Principal Component Analysis
UVE-DPLS	Discriminant variant of uninformative variable elimination partial least squares

4. BACKGROUND

The import and export of wines from the European Union is controlled in many ways. Regulations in respect to permissible wine manufacturing processes and ingredients are often complex. The policing of these regulations is largely undertaken by wine control laboratories in the importing and exporting countries. A multitude of tests are undertaken to determine the safety, authenticity and regulatory compliance of wines under test. Underpinning many of these analyses is a detailed knowledge of the processes that effect wine composition. This knowledge is supported by the utilisation of local, national or pan-European databases or databanks. These data can be used to confirm both the authenticity and geographical provenance of wine. However, to date there has been little consolidated effort to determine parameters that can be used for the determination of varietal purity or vintage in relation to wine. Both of these factors are of fundamental importance to the wine industry and the consumer as they can influence the value of wine and, being prominent in promotional labelling, are targets for falsified claims.

Many studies can be found in the literature where physico-chemical parameters of wines have been used to trace their geographical origin (e.g. country, region, province, and production area) [1], to study the impact of different grape varieties [2][3] and ageing processes [4]. However, the authors focus on a very limited number of parameters selected according to some preliminary knowledge. For instance, rare earth elements and isotopes can potentially be good indicators for tracing the geographical origin of wine thanks to the selective absorption phenomena [5].

Previous studies have highlighted the use of isotopic ratios [1][6][7][8][9], trace element composition [1][7][10][11], phenolic [12][13][14] and volatile profiles [15][16] in determining the geographical origin of wine. However, specific geographical markers are rarely that, and they are impacted to a greater or lesser extent by local practises and climate [7][8][11][12][13][16]. Many studies using geological markers as indicators of geographical origin are hindered by the complexity of their measurement and the temporal nature of elemental composition as an indicator. Indeed a quote from Suhaj & Koreňovská [11] when reviewing the use of trace element profiling for the determination of wine origin highlights this problem.

“Several factors, such as environmental contamination, agricultural practices, climatic changes, and vinification processes may change markedly the multi-element composition of the wine and may endanger the relationship between wine and soil composition”.

Climate also impacts on the ratio of light isotopes in wine and other commodities and therefore these measurements are not reliable long-term indicators of geographical origin. This is highlighted by Martin *et al*. [7].

“SNIF-NMR (Site-specific Natural Isotope Fractionation studied by NMR) and IRMS (Isotope Ratio Mass Spectrometry) parameters, which are subject to climatic influences, are usually more efficient for characterizing the year of production of a given appellation than for distinguishing different appellations”.

Varietal purity, growing practices and production processes have a significant impact on the phenolic and volatile composition of wine [12][13][16]. It is therefore clear that a robust system for the determination of geographical origin based on regional similarities may be an over optimistic goal using a single measurement system.

Compositional variation can provide an excellent indication of wine origin, but only when origin is considered with a higher degree of specificity than simple geographical location. Using specific matching techniques it is thus possible to take the analytical profile of a wine (all of the analytical measurements taken together) and determine the likely region of origin [1][17][18] [19] along with a range of other labelling claims.

Chemometric methods are extensively used to combine and simplify compositional data in relation to the verification of food authenticity. They support the process of experimental planning and the design of a sampling campaign, and are used to visualize and model the available multidimensional data. Specifically in food chemistry, many problems are of a discriminant nature and require the construction of multivariate discriminant models. The multivariate discriminant models are highly valued, since one parameter cannot provide sufficient information for the discriminant problem at hand. Properly validated discriminant models can help to distinguish between groups of samples of, for example different geographical origin, on the basis of their chemical characteristics.

The objective of this project was to study the possibility of discriminating wines prepared from various grapes grown in different years on the basis of a range of chemical parameters measured from the wines.

The goal of the European-founded project “Establishing of a WINE Data Bank for analytical parameters for wines from Third countries” (WINE-DB project, G6RD-CT-2001-00646-WINE-DB) [20] was to collect analytical data describing wine samples from different countries (Hungary, Czech Republic, Romania, South Africa and Australia) and to enable their successful discrimination according to geographical origin using chemometric methods. To date, results related to identification of geographical origin of wine samples have been presented in a series of articles [1][17][18][19]. In this study, we focus on the identification of parameters that can support discrimination of wines with respect to grape varieties and vintages. A total of 83 parameters were measured among which were the trace, rare earth elements and isotopes, organic and aroma compounds. Four discriminant techniques were applied, including discriminant partial least squares

(DPLS), the discriminant variant of uninformative variable elimination partial least squares (UVE-DPLS), classification and regression trees (CART) and neuro-fuzzy systems (NFS). These methods were particularly chosen since they help in handling linear and non-linear discriminant problems and in pointing out certain parameters that are potentially important for a given discriminant problem.

5. AIMS AND OBJECTIVES

The UK is the world's second largest importer of wine and one of the top ranking consumers with a wine import market in the region of £2bn. The UK currently lacks a system for the evaluation of wine origin using analytical methods. Furthermore, the controlling authorities in Europe have recorded a steadily growing number of mislabelled wine products with respect to their origin, vintage or other quality factors. The violation of the wine specification via, for example, unauthorized fortification and sugar addition are provoked by increasing competition for financial gain in the wine market.

A growing concern over consumers' health and satisfaction stimulates interest in the development of fast and cost-effective methods that can provide information about the origin of food (including beverage) samples (country, region or area), possible fraud or adulteration problems and about some aspects of food processing such as ageing and storage. These issues may greatly affect the economy of the food industry and therefore, many efforts are undertaken to identify unique parameters that can indicate potential problems. In agreement with this, one of the priorities of the European Community is to create regulations protecting the rights of both consumers and producers.

Objectives:

1) Construction of information portal: Construct a custom built searchable information portal to store all data in a standardised format.

2) Application of modelling techniques: Compare various models to determine which is the most appropriate to predict vintages and grape varieties of the experimental wines. This objective can be split into three sections:

   (a) Pre-processing of data.
   (b) Application of discriminant models to wines from Romania (proof of principle).
   (c) Application of models to the wines from the remaining countries.

3) Investigation of the use of other variable selection procedures. Examine the performance of univariate and multivariate variable selection approaches in order to optimise the method to discriminate vintage.

4) Other European countries (e.g. Germany) possess facilities for the analysis of both native and imported wines to perform a subjective determination of wine origin. An additional objective of this project is to report on methods used elsewhere.

6. OVERVIEW OF EXPERIMENTAL PROCEDURES

Experimental data

The data used in this project was originally recorded from sampling studies conducted on wines during the EU/FSA Third countries wine project (FSA project code: Q01063) to ensure data quality and wine integrity. This project concentrated on identification of geographical origin of wine samples. Authentic and commercial wines were included. The data from the Third Countries wine database was first checked for completeness and appropriateness for this project. Groups of grapes with more than 20 samples were considered and the following models were constructed:

a) vintages

- Authentic, red + whites from Romania - vintages 2002, 2003 and 2004,
- Authentic, red + white samples from Hungary - vintages 2002, 2003 and 2004,
- Authentic, red + white samples from Czech Republic - vintages 2002, 2003 and 2004,
- Authentic, red + white samples from South Africa vintages 2003, 2004 and 2005.

b) grape varieties

- Samples from Romania (authentic, red + white samples for all vintages): *Feteasca regala*.
- Samples from Hungary (authentic, red + white samples for all vintages): *Kekfrankos*, *Olasz Rizling*.
- Samples from Czech Republic (authentic, red + white samples for all vintages): *Riesling Italian*, *Muller Thurgau*, *Veltliner*.
- Samples from South Africa (authentic, red + white samples for all vintages): *Chardonnay*, *Cabernet Sauvignon*, *Chenin blanc*.

The original data contained 1308 samples and 105 parameters. Parameters that contain more than 60% missing elements were removed. Additionally the parameters describing the stable isotopes laboratory sample code and ethanol $\delta^{13}C$ were also excluded because of their discrete structure.

Twenty eight of the samples contained over 40% of missing elements and together with another four also containing incompleteness, were removed from the data. Thus, a total of 32 samples were removed. Since the rare earth ratios have been extensively studied [20], they were also included in the further analysis. These were the following ratios: Gd/La, Er/La, Yb/La, Gd/Er, Er/Yb. The final dataset contained 1274 samples and 83 parameters, containing 640 authentic samples and 634 commercial samples. Wines termed as authentic were manufactured specifically for analysis. The eighty three parameters included for each sample are detailed in Table 1. Statistical models were applied to the data in order to assess the best means (in terms of sensitivity and specificity) to discriminate the wines from one another.

Objective 1) Construction of information portal

A searchable and secure information portal was constructed early in this project. Wine analytical data and associated metadata was then uploaded to the portal and formatted so that data fields could then be filtered as required. Data was then exported directly from the portal to MATLAB (The Mathworks) for statistical modelling.

Objective 2a. Application of modelling techniques: pre-processing of data

All data parameters were log-transformed except: Malic acid, Na_excess, Ethanol $(D/H)_I$, Ethanol $(D/H)_{II}$, Ethanol $\delta^{13}C$ and Wine$\delta^{18}O$, which were left raw, and invert sugar was transformed using the following expression log(log(x))+2, where x is the variable of invert sugar. Additionally, depending on the chemometric method, a suitable pre-processing procedure was used.

Before presenting any discriminant model, exploratory data analysis using principal components analysis (PCA) was undertaken. PCA analysis is used to see the similarities of various wines with respect to grape variety and check for multivariate outliers, which were removed prior to further analysis. A detailed description of the analysis is given in Annex 1.

Table 1. A detailed list of all parameters measured in wine samples

No.	Name	No.	Name	No.	Name
1	Ethanolamine	29	Bromine	57	Ethanol $(D/H)_I$
2	Putrescine	30	Rubidium	58	Ethanol $(D/H)_{II}$
3	Ethylamine	31	Strontium	59	Alcohol
4	Lithium	32	Yttrium	60	Volume mass at 20°C
5	Beryllium	33	Molybdenum	61	Invert sugar
6	Boron	34	Cadmium	62	Volatile acidity
7	Sodium	35	Antimony	63	Tartaric acid
8	Magnesium	36	Cesium	64	D-lactic acid
9	Aluminium	37	Barium	65	L-lactic acid
10	Silicon	38	Lanthanum	66	Malic acid
11	Phosphorus	39	Cerium	67	Glycerol
12	Sulphur	40	Praseodymium	B	butanediol
13	Chlorine	41	Neodymium	69	Gluconic acid
14	Potassium	42	Samarium	70	Shikimic acid
15	Calcium	43	Europium	71	Methanol
16	Scandium	44	Gadolinium	72	Ethylacetate
17	Titanium	45	Terbium	73	1-propanol
18	Vanadium	46	Dysprosium	74	2-methyl-1- propanol
19	Chromium	47	Holmium	75	2-methylbutan-1-ol
20	Manganese	48	Erbium	76	3-methylbutan-1-ol
21	Iron	49	Thulium	77	Total alcohol
22	Cobalt	50	Ytterbium	78	Gd/La
23	Nickel	51	Lutetium	79	Er/La
24	Copper	52	Rrhenium	80	Yb/La
25	Zinc	53	Thallium	81	Gd/Er
26	Gallium	54	Lead	82	Er/Yb
27	Arsenic	55	Thorium	83	Wine $\delta^{18}O$
28	Selenium	56	Uranium		

The models constructed were:

Robust Principal Component Analysis (R-PCA)

Discriminant Partial Least Squares (DPLS)

Classification and Regression Trees (CART)

Kernel Discriminant Partial Least Squares (K-DPLS)

Discriminant Partial Least Squares extended with Uninformative Variable Elimination (UVE-DPLS)

Neuro-Fuzzy Systems (NFS)

Detailed information regarding the methods used and the reasoning behind the selection of each approach is included in Annex 1.

Objective 2b. Application of discriminant models to wines from Romania.

Initially, only data from authentic wines were considered and the discriminant models were built for wines from Romania in order to demonstrate the strategy of the analysis.

In order to verify that a discriminant model was able to distinguish between different vintages and grape type for Romanian wines, CART models were constructed and validated.

The aim of model construction is two-fold:

- to examine the type of models that provide best discrimination results (linear or non-linear models).
- to provide the smallest set of variables that are discriminative for the problem studied.

The results obtained for authentic Romanian red and white wines are presented in Annex 1, where sensitivities and specificities are obtained from the following type of models, constructed using the ‘one versus all’ principle:

- model 1 - 2002 vs. 2003 + 2004,
- model 2 - 2003 vs. 2002 + 2004
- model 3 - 2002 vs. 2003.

The analysis was performed separately for:

- red + white wines from Romania
- white wines from Romania.

In order to judge and compare the predictive power of models, the Monte-Carlo approach was considered. At each stage, the data were split randomly into model and test sets, such that the model set contained an equal number of objects from each group (circa 75% of the total number of samples in the smallest group). The model set was then used to build a discriminant model and the test set was used to evaluate its predictive power. The whole procedure was repeated *k* times and the final sensitivity and specificity were reported as averages over *k* repetitions. This approach is further detailed in Annex 1. The final sensitivity and specificity of a given type of model were reported as mean and median of the sensitivities and specificities obtained from 500 repetitions.

Objective 2c. Application of modelling techniques to wines from remaining countries

The supervised models described above were then built and validated for all remaining wines (Czech Republic, Hungary and South Africa). Again, to examine which models, e.g. linear or non-linear, gave the best predictions with a minimum set of variables, the Monte-Carlo approach was used. Further details are included in Annex 2.

Objective 3. Investigation of the use of other variable selection procedures

Two data sets were used to compare univariate versus multivariate variable selection, namely red and white wines from Hungary (45 samples of red wines and 99 samples of white wines) and the Czech Republic (38 samples of red wines and 113 samples of white wines) from vintages in 2002, 2003 and 2004. The performance of univariate variable selection procedures (with and without corrections for multiple testing) was evaluated. Among these approaches there were: the significance analysis approach (SA), t-test with two significance levels (α = 0.05 and 0.01), and t-test with two significance levels (α = 0.05 and 0.01) with the multiple hypothesis testing adjustments (Bonferroni [21], Holm [22] and Benjamini [23] corrections). The selected variables were then used to construct the DPLS model. Sensitivities and specificities for each model were reported. Vintage was then discriminated using univariate variable selection strategies prior to multivariate analysis. This was compared to the performance of multivariate variable selection methods.

7. OVERVIEW OF RESULTS

Objective 2a. Pre-processing of data

Initial PCA analysis of all wine samples identified one sample (South African wine sample) as an extreme outlier which was removed from the analysis.

Objective 2b. Discrimination of vintage and grape type of Romanian wines

It is possible to construct discriminant models capable of distinguishing different wine vintages and CART analysis gave a strong tendency for discriminating vintages. For example, the 2002 Romanian vintage showed low levels of arsenic while grapes harvested in 2003 showed the highest levels of arsenic, barium and wine $\delta^{18}O$.

Construction of separate models for all authentic wines (red and white wines together) and for white wines only, did not significantly improve discrimination. In general, the models constructed for all variables perform better in comparison with the models built for a limited number of variables. In some cases, the initial CART models can be further improved taking into account the linear combinations of selected variables. Constructing D-PLS models and non-linear models (K-D-PLS and NFS) did not significantly improve the discrimination results. Construction of discrimination models capable of distinguishing *Feteasca alba* from the other types was impossible. A more detailed account of the results is given in Annex 1.

Objective 2c. Discrimination of vintage and grape type of South African, Hungarian and Czech wines

(i) **Discrimination of vintage**

CART analysis was able to discriminate vintages. For example, the 2002 Romanian vintage showed low levels of arsenic while grapes harvested in 2003 showed the highest levels of arsenic, barium and wine $\delta^{18}O$. Other important variables for discriminating the grape variety were yttrium, glycerol, molybdenum, lanthanum, invert sugar and butadienol content. Depending on the vintage, eight to eighteen variables were required for discrimination.

It can be concluded that wines from a given country can be discriminated according to the year of production. All discriminant models are characterised by high sensitivities and specificities. In most cases, the use of a non-linear method (NFS-CART) improves the sensitivities and specificities of the models.

Vintage was discriminated for each country with the best results for wines from the Czech Republic and in general all models showed high sensitivities and specificities, with NFS-CART tending to give the best sensitivity and specificity. The difference between the wines is emphasised by the fact that different variables are used to discriminate vintage for each country of origin (Table 2, below). The sensitivities and specificities of each model used are detailed in Tables 2b, c, d and e of Annex 2). A different set of variables is important for each individual model and a total of 43 variables should be measured to predict vintage (Table 3, below). A more detailed account of the results investigating the discrimination of vintage and grape type of South African, Hungarian and Czech wines is given in Annex 2.

Table 2. Summary of the 5-6 variables used in D-PLS-CART, kernel-PLS-CART and NFS-CART to distinguish vintage for each of the countries of origin.

Country / Vintage	Czech Republic	Hungary	South Africa	Romania
2002	Cd, Ti, S, B, Ca, Be	S, Na, Er/La, invert sugar, ethylacetate, Gd/La	As, ethanolamine, Fe, methanol, Sb, Ba	Yb, Sb, Be, As, putrescine, Tm
2003	Wine $\delta^{18}O$, Mg, Al, Si, U, ethanolamine	Wine $\delta^{18}O$, S, Ti, Ca, Sc, Re	Be, As, Yb/La, ethanolamine, Fe, Mo	Wine $\delta^{18}O$, Ba, Be, Mn, Se, glycerol
2004	Cr, S, U, Sc, Be, ethanolamine	Wine $\delta^{18}O$, alcohol, malic acid, Cr, Ti, S	Fe, Se, Be, Ba, ethanolamine, Mo	Ethanolamine, Br, Be, Ba, Se, butadienol

(ii) Discrimination of grape variety

The sensitivity and specificity of different models were compared for the discrimination of grape variety (Table 4, Annex 2). In general the best discrimination method was NFS-CART. This is advantageous since CART models are preferred over other methods such as UVE-PLS because of their specificity combined with their simplicity.

Wines of a given origin can be distinguished according to grape variety by this method. High sensitivities and specificities are seen for the South African *Cabernet Sauvignon* and the Hungarian wines from the *Kekfrankos* grape (highest sensitivities 99.63% and 97.5% respectively). The Italian

Reisling and Czech *Veltliner* grapes were most difficult to discriminate (with highest sensitivities of 86.91% and 72.75% respectively). A total of 36 variables were required to characterize grape variety as shown in Table 4. It should be noted that the Romanian grape *Feteasca alba* could not be distinguished from the other grapes.

Table 3. Variables selected by CART to discriminate according to vintage regardless of geographic origin.

No.	Index of variable	Variable name		No.	Index of variable	Variable name
1	1	Ethanolamine		23	35	Sb
2	2	Putrescine		24	37	Ba
3	3	Ethylamine		25	38	La
4	5	Be		26	49	Tm
5	6	B		27	50	Yb
6	7	Na		28	52	Re
7	8	Mg		29	55	Th
8	9	Al		30	56	U
9	10	Si		31	59	Alcohol
10	11	P		32	60	Volume mass at 20ºC
11	12	S		33	61	Invert sugar
12	15	Cl		34	62	Volatile acidity
13	16	Sc		35	66	Malic acid
14	17	Ti		36	67	Glycerol
15	19	Cr		37	68	Butanediol
16	20	Mn		38	71	Methanol
17	21	Fe		39	72	Ethylacetate
18	27	Co		40	78	Gd/La
19	28	Se		41	79	Er/La
20	29	Br		42	80	Yb/La
21	33	Mo		43	83	Wine $\delta^{18}O$
22	34	Cd				

For the Czech wines, using 6-13 variables for each vintage, models were built using CART-D-PLS or UVE-PLS with a sensitivity of ≥99.31%. Further, using univariate selection approaches, sensitivities of >98% were seen with different variables being important for different vintages. For the wines originating from Hungary, constructing models from subsets of variables did not improve the sensitivity of the models. In terms of increasing the simplicity of the models and in order to minimise the occurrence of false positives by reducing the number of variables, models could be constructed with sensitivities and specificities >90% for a very limited number (5-6) of variables. A more detailed account of the results is included in Annex 2.

Objective 3. Investigation of the use of other variable selection procedures

(i) Discriminant models for wines from the Czech Republic

For the Czech samples, in the case of models to discriminate the 2002 vintage, a relatively satisfactory model was obtained using the (linear) CART approach, offering 97.12% and 96.99% of sensitivity and specificity, respectively. Taking into account only six explanatory variables, selected using CART on the basis of their selection frequency in 500 CART models, the D-PLS model constructed obtained 100% sensitivity and 99.99% specificity. Therefore, it can be considered as the best one and no further discrimination improvement was necessary.

Table 4. All important variables selected by CART to discriminate wines according to grape variety regardless their geographical origin.

No.	Index of variable	Variable name
1	1	Ethanolamine
2	2	Putrescine
3	3	Ethylamine
4	4	Li
5	5	Be
6	6	B
7	7	Na
8	8	Mg
9	10	Si
10	11	P
11	12	S
12	13	Cl
13	14	K
14	15	Ca
15	18	V
16	21	Fe
17	22	Co
18	25	Zn
19	26	Ga
20	27	As
21	30	Rb
22	34	Cd
23	38	La
24	52	Re
25	53	Tl
26	58	Ethanol $(D/H)_{II}$
27	62	Volatile acidity
28	63	Tartaric acid
29	64	D-lactic acid
30	70	Shikimic acid
31	71	Methanol
32	73	1-propanol
33	74	2-methyl-propanol
34	75	2-methylbutan-1-ol
35	76	3-methylbutan-1-ol
36	81	Gd/Er

From all 2003 models, the (linear) D-PLS-CART model with six explanatory variables selected with the CART approach can be considered as the best in terms of its sensitivity (99.72%), specificity (99.32%), and the number of variables. From 2004 models, the best discriminant model was constructed using the UVE-D-PLS approach and 13 variables. Sensitivity and specificity of the final D-PLS model built for 13 variables are equal to 99.31% and 97.79%, respectively.

(ii) Discriminant models for wine samples from Hungary

In the case of 2002 models the best discriminant model (in terms of sensitivity and specificity) was constructed using the D-PLS approach for all explanatory variables. This model is characterized by sensitivity and specificity equal to 91.73% and 92.83%, respectively. A slightly worse discrimination result was achieved for a subset of six explanatory variables, pre-selected using the CART strategy, for which the D-PLS model has sensitivity and specificity of 89.96% and 90.83%, respectively.

For 2003 models, as shown in Table 5, Annex 3, very good sensitivities (above 95%) and specificities (above 92%) were obtained from the D-PLS model constructed for variables selected using CART (six variables) and the UVE-D-PLS approach (ten variables). Taking into account non-linear models, constructed either for all explanatory variables or their subset, does not improve overall discrimination results.

From 2004 models, the highest sensitivity (97.54%) and specificity (94.91%) are obtained from the D-PLS model, constructed from all explanatory variables (see Table 5, Annex 3). Worse discrimination performance (sensitivity equal to 93.14%, and specificity 93.25%) is obtained from the D-PLS model, constructed for six variables selected using CART. The variables used to distinguish vintage in Czech Republic and Hungary are shown in Table 2, Annex 3. Full details of the results for this objective are included in Annex 3.

8. SUMMARY OF WINE CONTROL PROCEDURES IN GERMANY

Carsten Fauhl-Hassek, Head of the Senior Expert Office for the Import Control of Wine, Federal Institute for Risk Assessment, BfR, Berlin, Germany, has provided a detailed report of wine control practise in Germany which is included in Annex 4.

The Official Wine Control laboratories in Germany have much experience in the authenticity of wine, testing national, European and international products for conformity with regulation. BfR is the senior expert office for the import control of Third country wines and, in this function, manages a range of databanks of the official national wine bodies, stocked with reference data. Information regarding the analytical aspects relevant to wine control is summarised below.

The major aims of Official Wine Control in Germany are:

- control of sensory quality
- labelling control (including alcohol content, quality, grape(s), geographic origin, vintage etc. and a so-called '85% rule' applies regarding the mixing of wines. Determining the geographic origin of a wine is the most challenging aspect
- determination of chemical adulterations (including water, glycerol, alcohol, sugar, dyes, sweeteners, preservatives and flavourings)

Applicable parameters include:

- Preservatives (sorbic acid, benzoic acid)
- Dyes
- Mineral composition (Na, K, Ca, Mg, ash)
- Anions (Cl^-, SO_4^{2-}, PO_4^{3-}, NO_3^{2-})
- Glucose/fructose
- Acids (tartaric acid, malic acid, lactic acid, glucoronic acid)
- Glycerol, butandiol, higher alcohols
- Glycerol by-products
- Anthocyan ratios
- Shikimic acid
- Biogenic amines (often tested in case of consumer complaint or focussed surveys)

- Heavy metals, pesticides, mycotoxins (ochratoxin) parameters tested in focussed surveys, not normally routine

Specialised laboratories also test for:

- Stable isotope ratios (D/H, $^{13}C/^{12}C$, $^{18}O/^{16}O$)
- Flavour analysis (e.g. pyrazines)
- Enantioselective analysis (e.g. enatiomeric ratio analysis of lactones)

In terms of control limits, some are set down by European law, some by national law or recommendations by the OIV (International Organisation of Vine and Wine). Factors which pose a risk to human health have control limits. For other parameters there are no control limits e.g. stable isotopes which are indicative of geographic origin/climate conditions.

The content levels for certain parameters can be indicative of adulteration of wines:

- High or low glycerol factor can be indicative of addition of glycerol, ethanol or water.
- Mg^{2+} <50mg/L can be indicative of water addition
- NO^{3-} >30mg/L can be indicative of water addition
- Na^{+} >80mg/L can be indicative of de-acidification with NaOH or application of ion exchange
- Low Ca^{2+} is indicative of unauthorised procedures

Stable isotope analysis has proved a valuable form of evidence in terms of judicial cases:

- ^{2}H indicative of chaptalisation. Method has not been validated for wine, although has been for fruit juice, so there are no precise guidelines.
- $^{18}O/ ^{16}O$ can be indicative of watering wine when compared to authentic wine grown in the same region and climate. Official OIV and EEC methods exist but no limits stated.
- $^{13}C/ ^{12}C$ can be indicative of addition of alcohol or cane sugar. OIV method exists.
- $^{13}C/ ^{12}C$ and $^{18}O/ ^{16}O$ give information regarding growing climate/origin.

However, there is no limit for some minerals and decisions lie with experienced analysts to decide whether adulteration may have taken place. Analysts rely heavily on reference databanks and must take consideration of measurement uncertainty, by comparing results to those obtained from at least 30 reference samples and with reference to relevant statistics. Measurement uncertainty must also be considered.

Due to the potentially large number of parameters to be tested, the testing laboratory must find a compromise between the number of samples tested and the targeted control of wines of especially suspicious origin in terms of cost and time spent.

The official German wine control has constructed their own $\delta^{18}O$/ $\delta^{16}O$ databank to check origin and verify watering of wines. The databank, which is updated bi-annually, contains around a thousand entries from South America, USA, North and South Africa, Australia, New Zealand and Eastern Europe and wines are reported according to country and vintage. Mean values and standard deviations are used to determine upper and lower limits with a 95% confidence. However, there is a need to take account of extreme weather or differences between growing regions and to consider all of the information available. Vintage fluctuations play an important role in certain regions e.g. mid-Europe, and labelling information regarding origin may be incorrect. Both factors could skew the data, especially with the 85% mixing rule.

The concentration of shikimic acid is indicative of grape variety and quantitative liquid chromatography methods have been validated. BfR hold a databank of this information also but again the 85% rule adds variation to the shikimic acid distribution so it becomes impossible to predict the grape variety. There are no official cut-offs for stable isotope ratios or for many minerals in wines so databanks are an important point of reference.

Anthocyan chemistry and ratio measurements are indicative of grape variety. This method has been adopted by the OIV but is complicated by the fact that the extraction method is affected by the fermentation temperature/duration and concentrations of sulphur dioxide and alcohol.
Glycerol is a natural product of fermentation with around 92% of sugar converted to alcohol and 8% converted to glycerol. Addition of glycerol is fraudulent as its presence is linked to the good mouth feel of quality wines and methods exist for the quantitative determination of glycerol and alcohol. However, since natural variation hinders the reliable determination of fraud, the addition of industrial/synthetic glycerol is determined by detecting industrial impurities by gas chromatography mass spectrometry. This is an OIV-validated method although variation exists between sources of glycerol. The German wine control have determined enforcement limits of 0.1mg/L for 3-methoxy-

1,2-propanediol (3-MPD) and 0.5mg/L for cyclic diglycerols (CycDs), both industrial impurities of glycerol. Pyrazines also exist naturally in wines and contribute to the sensory quality. Natural levels vary but should not exceed 40ng/L.

In conclusion, wine quality control is a complex science which requires much experience and detailed reference databanks due in part to factors such as natural variations in parameters and legitimate mixing of wines. Accreditation in the techniques involved is necessary to reliably determine the quality of the various parameters measured in wines. Dedicated facilities are recommended for wine analysis with analysts trained comprehensively by an experienced control laboratory. An important point to note is that authenticity testing is criticised as it does not always sufficiently take account of natural diversity, biological or oenological effects so building databases over time helps to overcome this. Dealing with authentic samples is therefore linked to a certain level of probability and false negatives or false positives.

9. WEB-BASED PORTAL

A searchable and secure information portal was constructed early in this project. All of the wine analytical data and associated metadata available (1308 samples and 105 parameters) from the Third Countries Wine Project was uploaded to the portal and formatted so that data fields could then be filtered as required. Search tools and output facilities were built in to enable data handling. Data was then exported directly from the portal to MATLAB (The MathWorks) for pre-processing and statistical modelling.

10. CONCLUSIONS

In the present study, wines of a given origin could be distinguished according to grape variety with high sensitivities and specificities in most cases. A sensitivity of 99.6% for the South African *Cabernet Sauvignon* and 97.5% for Hungarian wines from the *Kekfrankos* grape was seen by DPLS-CART. Non-linear models perform better than linear models. The poorest models were obtained for the *Riesling* Italian (66.2%) and *Veltliner* grapes (70.0%) from the Czech Republic.

Wines could also be discriminated according to vintage and, in general, all models gave high sensitivities and specificities. In most cases, the use of a (non-linear) NFS-CART method using five variables improved the sensitivities and specificities of the models. Higher numbers of variables (e.g. six) would increase sensitivity and specificity but would require a (perhaps prohibitively) long computation time. The variables selected were different for each country of origin and for each year of harvest which emphasises the difference between the wines. A combination of up to forty three variables is needed to predict one grape from the selection of nine grapes in this survey.

On the basis of samples collected in the Czech Republic and Hungary, it is possible to build satisfactory D-PLS models using only a limited number of variables to discriminate vintage. From a statistical point of view, variable selection approaches minimising the potential for false positives should be preferred, although in only a few cases did multivariate analyses improve the performance of univariate models.

As highlighted in the current study, a number of different measurements must be taken for each suspect wine in order to evaluate authenticity. Although six variables could be used in this study to obtain a satisfactory CART model, the project concludes that the recommended method for determining the authenticity of the wines involves measurement of a total of 43 variables to predict vintage and 36 to predict grape variety of a new wine sample. This is commensurate with German control practice where a large number of variables are investigated for a wine under investigation and labelling information is used to aid in the decision of parameter selection.

Future work is recommended to concentrate on further developing a databank of wines from countries and grapes of import interest to the UK. Wines should be tested regularly and data uploaded to the database as part of an ongoing process to gain experience of natural variations in natural diversity, biological or oenological affects. The development of databanks and further experience in this area seems paramount to the initiation of a UK wine testing facility.

A report from one other EU country (Germany) was sourced during the time frame of this study and is included in Annex 4. A second report from another EU wine-producing country will be provided at a later date.

It is important that the information gained from this project is disseminated to the relevant scientific and wine-testing communities. We propose a web-based learning package to include on-line presentations which could be developed at Fera, that interested parties could access at any time. The learning package would be advertised to potentially interested parties by email.

11. ACKNOWLEDGEMENTS

We gratefully acknowledge the Food Standards Agency, UK for funding this study. We also thank Prof. Baeta Walczak, Michal Wrobel, Dr. Michal. Daszykowski and Ivana Stanimirova for their support in performing the statistical analysis of the data. We are also grateful to Dr. Carsten Fauhl-Hassek for providing detailed information regarding wine control practise in Germany. Further, we are grateful to the EU project partners for allowing the use of the data generated from the EU/FSA project 'establishing of a wine databank for analytical parameters for wines from Third Countries'(G6RD-CT-2001-00646-WINE-DB). Additionally we thank Dr. C. Guillou (European Joint Research Centre, Ispra, Italy), Dr. B. Medina (Directeur du Laboratoire de Bordeaux/Talence, and Prof. R. Wittkowski (BfR, Germany), Dr. Mark Woolfe (Food Standards Agency, UK) and Anaisabel Blanch (Ministerio de medio Ambiente y medio Ruraly Marino, Spain) for their valued participation in the kick-off meeting to this project, which was also attended by Prof Walczak and Dr Fuhl-Hassek.

12. REFERENCES

[1] X. Capron, J. Smeyers-Verbeke, D. Massart, Food Chemistry, 101 (2007) 1585-1597.

[2] J. Câmara, M. Alves, J. Marques, Talanta, 68 (2006) 1512-1521.

[3] S. Roussel, V. Bellon-Maurel, J. Roger, P. Grenier, Journal of Food Engineering, 60 (2003) 407-419.

[4] in: Comprehensive Chemometrics, Elsevier, Amsterdam, 2009, pp. 75-128.

[5] M. Forina, C. Armanino, M. Castino, M. Ubigli, Vitis, 25 (1986) 189-201.

[6] P. Coetzee, F. Vanhaecke, Analytical and Bioanalytical Chemistry, 383 (2005) 977-984.

[7] G. Martin, M. Mazure, C. Jouitteau, Y. Martin, L. Aguile, P. Allain, American Journal of Enology and Viticulture, 50 (1999) 409-417.

[8] J. Giménez-Miralles, D. Salazar, I. Solana, Journal of Agricultural and Food Chemistry, 47 (1999) 2645-2652.

[9] M. Barbaste, K. Robinson, S. Guilfoyle, B. Medina, R. Lobinski, Journal of Analytical Atomic Spectrometry, 17 (2002) 135-137.

[10] M. Baxter, H. Crews, M. Dennis, I. Goodall, D. Anderson, Food Chemistry, 60 (1997) 443-450.

[11] M. Suhaj, M. Koreňovská, Acta Alimentaria, 34 (2005) 393-401.

[12] J. Cortell, M. Halbleib, A. Gallagher, T. Righetti, J. Kennedy, Journal of Agricultural and Food Chemistry, 53 (2005) 5798-5808.

[13] K. Russell, S. Zivanovic, W. Morris, M. Penfield, J. Weiss, Journal of Food Quality, 28 (2005) 377-385.

[14] D. Picque, T. Cattenoz, G. Corrieu, J. Berger, Sciences des Aliments, 25 (2005) 207-220.

[15] M. Martí, O. Busto, J. Guasch, Journal of Chromatography A, 1057 (2004) 211-217.

[16] U. Fischer, D. Roth, M. Christmann, Food Quality and Preference, 10 (1999) 281-288.

[17] J. Smeyers-Verbeke, H. Jäger, S. Lanteri, P. Brereton, E. Jamin, C. Fauhl-Hassek, M. Forina, U. Römisch, European Food Research and Technology, 230 (2009) 15-29.

[18] U. Römisch, H. Jäger, X. Capron, S. Lanteri, M. Forina, J. Smeyers-Verbeke, European Food Research and Technology, 230 (2009) 31-45.

[19] M. Forina, P. Oliveri, H. Jäger, U. Römisch, J. Smeyers-Verbeke, Chemometrics and Intelligent Laboratory Systems, 99 (2009) 127-137.

[20] Establishing of a wine data bank for analytical parameters for wines from third countris, Bundesinsitiut für Risikobewertung, De, March 2006.

[21] Bonferroni, C. E. Studi in Onore del Professore Salvatore Ortu Carboni. Rome: Italy, pp. 13-60, 1935 and Weisstein, E.M, "Bonferroni Correction" MathWorld - A Wolfram Web Resource; http://mathworld.wolfram.com/BonferroniCorrection.html.

[22] Holm, S. (1979) Scandinavian Journal of Statistics **6** (2): 65–70.

[23] Benjamini, Y and Hochberg, Y (1995) Journal of the Royal Statistical Society, Series B (Methodological) **57** (1): 125–133.