FOODINTEGRITY

Ensuring the Integrity of the European food chain

613688: Collaborative Project

Seventh Framework Programme
KBBE.2013.2.4-01: Assuring quality and authenticity in the food chain

Deliverable D18.4

Report on the validation and implementation of the untargeted protocol

Author(s): Emiliano De Dominicis, Marialuisa Piva, Elisa Gritti, Monica Locatelli, Marco Arlorio, Cristiano Garino, Maurizio Rinaldi, Luigi Portinale, Giorgio Leonardi

Beneficiary(s): P59

Date of preparation: 31/08/2018

Period covered: 01/06/2018 - 31/08/2018

Status: version 1

Dissemination level		
PU	Public	
PP	Restricted to other participants	X
RE	Restricted to a group specified by the consortium	
CO	Confidential, only members of the consortium	

The project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement No. 613688.

TABLE OF CONTENTS

Deliverable 18.4 Report on the validation and implementation of the untargeted protocol

1. Description of Deliverable

The main objective of this deliverable is to present a practical example of an *in house* (intra-laboratory) validation protocol for untargeted methods. Based on the different traceability/authentication cases considered in the WP18, we selected the identification of the geographic origin, specifically Italy vs East Europe, of Acacia honey, as the most suitable model for describing our suggested methodological approach. The selected method was performed using Liquid Chromatography coupled with Low Resolution Mass Spectrometry (LC-LRMS) technique and *multiclass (binary)* classification model (for more details about the selection of the method, see the Deliverable 18.3).

Following the activities described in the Deliverables 18.2 (Collection of all databases for each selected matrix) and 18.3 (Selection of the most performing method to be used as model for the validation), the realization of the present deliverable was obtained through the following actions:

- Definition of the best parameters of classification method and their calculation for the firstly collected data set of Acacia honey (internal validation);
- Collection and analysis of another set of Acacia honey samples, in order to complete the validation process by testing the prediction ability of the developed method on blind samples (samples unknown to the classifiers, but whose origin is known to the analyst) (external validation).

Due to the intrinsic limitations of the selected analytical technique (different laboratories and/or instruments do not usually obtain equivalent data sets that can be integrated in or

evaluated by the same developed method), in the present deliverable we have proposed an example of an *in house* validation protocol.
A general guideline about the validation approach proposed for this example will be the object of the final deliverable of WP18 "Good practices and methodological guidelines for the validation and application of the untargeted analysis for food authenticity and traceability" (Deliverable 18.5).

The deliverable is organized as follows: in *section 2* the main parameters suggested for the validation of untargeted methods are described; in *section 3* the example of validation protocol applied to the identification of geographic origin of Acacia honey is presented. In particular, *section 3* includes the purpose of the study (*section 3.1*), the description of the data set (*section 3.2*), the classification approaches (section 3.3) and the results of the internal (section 3.4) and external (section 3.5) validation.
A systematic discussion of the results, evidencing suggestions and/or criticisms, is reported in *section 4* (Achievement of the Deliverable).

2. Description of the validation parameters

This section presents the main validation parameters and measures for assessing the overall quality of the classifier models, in terms of their ability in recognizing the correct class labels assigned to the analysed food samples. These parameters rely on the definition of: ***positive samples*** *(**P***, samples belonging to one or more classes of interest) and ***negative samples*** *(**N,*** all the other samples). The attempt of one classifier to identify them, by assigning labels starting from the samples' features, generates 4 different results:

- **True Positives (TP):** the number of the positive samples that were correctly labelled by the classifier
- **True Negatives (TN):** the number of negative samples that were correctly labelled by the classifier
- **False Positives (FP):** the number of negative samples that were incorrectly labelled as positive
- **False Negatives (FN):** the number of positive samples that were mislabelled as negative

These terms are summarized in a confusion matrix, shown in Figure 1.

		Predicted class		
		yes	*no*	Total
Actual class	*yes*	*TP*	*FN*	*P*
	no	*FP*	*TN*	*N*
	Total	P'	N'	$P + N$

Figure 1. Confusion matrix, with totals for positive and negative samples

The confusion matrix is an important information about the validation, since it can show the picture about how many samples have been classified correctly (***TP*** out of ***P***, and ***TN*** out of ***N***) and how many have been misclassified (the remaining ***FN*** and ***FP***), for each class. The table may have additional rows or columns to provide totals. For example, in the confusion matrix of Figure 1, ***P*** and ***N*** are shown. In addition, ***P'*** is the number of samples that were labelled as positive (***TP*** + ***FP***) and ***N'*** is the number of samples that were labelled as negative (***TN*** + ***FN***). The total number of samples is ***TP*** + ***TN*** + ***FP*** + ***TN***, or ***P*** + ***N***, or ***P'*** + ***N'***. Note that, although the confusion matrix shown is for a binary classification problem, confusion matrices can be easily drawn for multiple classes in a similar manner.

On the basis of this matrix, a number of evaluation measures are obtained, and used as validation parameters for our goals:

- $\boldsymbol{Accuracy\ rate} = \frac{TP+TN}{P+N}$: reflects how well the classifier recognizes tuples of the various classes;

- $\boldsymbol{Kappa\ statistic\ (K)} = \frac{Pr(A)-Pr(e)}{1-Pr(e)}, Pr(a) = \frac{TP+TN}{TP+FP+TN+FN}, Pr(e) = \frac{P'*P+N'*N}{(P+N)^2}$: measures inter-rater agreement for qualitative (categorical) items, and can be seen as a measure for assessing accuracy and reliability of a statistical classification;

- $\boldsymbol{Precision} = \frac{TP}{TP+FP}$: measures the "*exactness*" of the prediction, in terms of what percentage of samples labelled as positive are actually such;

- $\boldsymbol{Sensitivity\ or\ Recall\ or\ Non-error\ rate} = \frac{TP}{P}$: also defined as *true positive rate*, measures the "*completeness*" of the prediction, which is the proportion of the positive samples correctly identified;

- $\boldsymbol{Specificity} = \frac{TN}{N}$: also defined as *true negative rate*, is the proportion of the negative samples correctly identified;

- $\boldsymbol{Fall\text{-}out} = \frac{FP}{N}$: also defined as false positive rate, is the proportion of negative examples incorrectly classified. It is equal to 1-specificity;

- ***ROC (Receiving Operating Characteristic) Curve***: is the plot of the *sensitivity* of a classifier as a function of the *fall-out* (1-*specificity*); in other terms is the curve obtained by plotting and joining a set of points *(fpr,tpr)* where *fpr* is the false positive rate and *tpr* is the true positive rate of a classifier. Different curves can be obtained by changing the classification parameters of the model (e.g., the threshold for classifying the positive class);

- $\boldsymbol{Area\ under\ ROC\ curve}$: it is the area of the region under the ROC curve; it can be interpreted as the probability that the model ranks a random positive example more highly than a random negative example. The closer to 1 is this value, the better the performance of the classifier is.

Accuracy rate, ***Precision*** and ***Sensitivity*** can be expressed in the range 0-1 or as percentage values (0-100%).

The following are aggregate measures, summarizing different validation parameters for a further inspection of the classifiers performances:

- $\boldsymbol{Matthews\ correlation\ coefficient\ (MCC) = \frac{TP*TN-FP*FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}}$: measures the correlation between the observed and predicted binary classifications;

- $\boldsymbol{F\text{-}measure\ (F_1) = 2\frac{precision*recall}{precision+recall}}$: represents the harmonic mean of precision and recall.

In case of class unbalance, which occurs when the classes of interest in a dataset are not equally represented, appropriate measures should be taken. In fraud detection, for example, the class, generally speaking, “fraud” occurs much less frequently than the generally defined “non-fraudulent” class. In these cases, if a multi-class classification approach is used, it is important to concentrate on validation parameters able to manage the unbalanced distribution of the samples, assessing how well the classifier can recognize the positive samples, and how well it can recognize the negative samples.

To achieve this goal, it is possible to concentrate (mainly) on sensitivity and specificity per class, and on the area under the ROC curve. Additionally, cost-based validation measures should be considered.

Cost-based evaluation relies on a cost matrix, assigning, for each class *i*, the cost of predicting a sample belonging to *i*, to the same class *i,* or to a different class $j \neq i$. A cost matrix has the following form:

		PREDICTED CLASS	
ACTUAL CLASS	C(i|j)	**Class=Yes**	**Class=No**
	Class=Yes	C(Yes|Yes)	C(No|Yes)
	Class=No	C(Yes|No)	C(No|No)

Figure 2. A cost matrix for binary classification

The costs can be assigned on the basis of the cardinality of each class, or exploiting knowledge about the application domain (assigning a higher cost to misclassification of the negative samples, for example) or combining both of these aspects.

Additionally, a-priori probabilities P_g can be set for each class *g*, in a uniform way ($P_g = 1/G$, for each class *g*, with G = no. of classes) or using the cardinalities of each class (e.g. $P_{yes} = P / (P+N)$; $P_{no} = N / (P+N)$).

After assigning costs and probabilities, the cost-based misclassification risk for each class *g* can be calculates as:

$$MR_g = \sum_g \frac{\left(\sum_{g'} CM_{gg'} CS_{gg'}\right) P_g}{N_g}$$

Where ***g*** is the considered class, $CM_{gg'}$ is the value of classifying ***g*** with ***g'*** in the confusion matrix, $CS_{gg'}$ is the corresponding cost in the cost matrix, P_g is the a-priori probability for class ***g*** and N_g is the number of samples in the class ***g***. The total misclassification risk is the sum of all the risks for each class.

Finally, a classifier assigns a class to each unknown sample on the basis of a probability value, which can be seen as a confidence value assigned by the classifier to the considered prediction. Therefore, another validation parameter which could be useful to considered is the mean and variance of the probabilities assigned to the samples classified as ***TP***, ***TN***, ***FP*** and ***FN***.

All of these validation parameters must be carefully chosen for both the internal and external validation, and will be produced for each classification model obtained by the elaboration of the honey samples analysed for the case study presented in this deliverable.

3. Description of the study (assessment of geographic origin of Acacia honey)

3.1 Purpose of the study and identification of the acceptance thresholds for the validation

The general purpose of the present study is to develop a preliminary non targeted method able to discriminate Italian Acacia honey from Eastern European one. More particularly, the specific aim is to test the potential application of LC-LRMS technique to this specific problem, thus confirming, as future perspective, the possibility to implement the method, making it a routine method for the quality control at industrial level.

Concerning the internal validation, the method will be considered acceptable for Accuracy, Sensitivity and Specificity values of at least 95%, above which the following step of external validation can be carried out.

In order to accept the overall method for future implementations (considering other variables as seasonality, raw materials, manufacturing process, packaging, transport, storage and so on), the external validation will not be considered acceptable for Accuracy and Sensitivity values lower than 75%.

The Misclassification Risk was determined by defining a cost matrix to measure only the impact of the misclassifications; therefore, we placed the cost 1 in the positions corresponding to misclassifications in the cost matrix (C(no|yes) and C(yes|no)), and the cost 0 elsewhere. Misclassifications of Italian honey samples were considered equivalently to misclassifications of East European ones. A-priori probabilities have been set considering the cardinality of the samples for each class. Therefore, the a-priori probabilities for the misclassification risk in the internal validation dataset are set as: $P_{Italian}$ = 90/207 and $P_{East\ Europe}$ = 117/207, while the a-priori probabilities for the external validation dataset are set as: $P_{Italian}$ = 60/246 and $P_{East\ Europe}$ = 186/246.

3.2 Description of the data set

3.2.1 Samples

Honey samples for both internal and external validation sets were provided by an Italian dealer that guaranteed the origin of honey samples. All the samples were provided in glass jars (closed with screw caps) and stored in the dark at room temperature until the sampling for the extraction and the analysis.

- *Internal validation set*

A first set of 69 Acacia honey samples was obtained for developing the analytical method and the first step of the validation process (internal validation). More particularly, 30 honey samples were from Italy and 39 from East Europe. Eastern European honey samples included

15 honey from Hungary, 3 from Romania, 3 from Serbia, 1 from Moldavia; the origin of the remaining honey samples was declared as not quantitatively defined mixtures of the previously cited East European countries.

- *External validation set*

A second set of 82 Acacia honey samples (20 from Italy, 62 from East Europe (including Hungary, Romania, Serbia, Moldavia, Croatia and Ukraine) was obtained to fulfil the second step of the validation process (external validation with blind samples). The specific composition of Eastern European samples was as follows: 2 from Croatia, 10 from Romania, 24 from Hungary; the remaining 26 samples were indicated as different mixtures from the previously specified countries.

3.2.2 Analytical method

Each sample was extracted in single and analysed in triplicate; analyses were performed by using a SCIEX QTRAP 6500+ LC-MS/MS system coupled with a 3000 Ultimate Thermo UPLC. One gram of honey was weighted in a plastic ultracentrifuge tube and added to 3 ml of water. After 8 min shaking, 6 ml of 1% CH_3COOH in CH_3CN were added, and the sample was stirred for 10 min. Phases separation was achieved by addition of 4 g of $MgSO_4$ and 1 g of sodium acetate. Samples were stirred for 10 min, then centrifuged and left in refrigerator for almost 2 hours. The acetonitrile phase of each sample was divided in two aliquots: one to be analysed in positive mode (POS) and one to be analysed in negative mode (NEG). The POS aliquot was purified with 300 mg of $MgSO_4$ and 75 mg of PSA sorbent bulk, centrifuged and diluted 1:1 with ammonium formate 10mM pH 4 containing internal standard Chlorpyrifos Ethyl D10. The NEG aliquot was purified with 200 mg of $MgSO_4$ and 35 mg of C18 sorbent bulk, centrifuged and diluted 1:1 with ammonium formate 10mM pH 4 containing internal standard Nicarbazine.

Operating conditions:

- ***LR-MS***

Scan Mode:	Full scan; scan range: 50-1000 Da
Polarity:	positive/negative
Polarity:	**POS**
Curtain gas (CUR):	30 psi
Collision Gas (CAD):	Medium
Ion Spray Voltage (IS):	+5500 V
Temperature (TEM):	450°C
Ion Source Gas 1 (GS1):	45 psi
Ion Source Gas 2 (GS2):	40 psi
Declustering Potential (DP):	80

Entrance Potential (EP): 10

Polarity: **NEG**
Curtain gas (CUR): 30 psi
Collision Gas (CAD): Medium
Ion Spray Voltage (IS): -4500 V
Temperature (TEM): 450°C
Ion Source Gas 1 (GS1): 45 psi
Ion Source Gas 2 (GS2): 40 psi
Declustering Potential (DP): -70
Entrance Potential (EP): -10

- ***LC***

1. Column: *Acquity UPLC BEH C18 130Å (50 mm x 2.1 mm i.d. x 1.7 µm)*
2. Flow: 0.2 ml/min
3. Mobile phase: A) Ammonium formate 10 mM pH4
   B) MetOH

Time (min)	%B
0,0	0
0,5	0
14,0	100
16,0	100
16,5	0
19,0	0

3.2.2.1 Column conditioning and LC-LRMS injection plan

The column was used and maintained according to the supplier's indications, and was checked twice before analysing the first sample.

In order to avoid systematic bias due to analytical variation, all samples were injected under a randomized sequence, and QC samples, obtained by mixing an equal amount of 15 Italian honey samples and 15 East European honey samples, were analysed at regular intervals (about every other 20 samples, corresponding to the start, the middle and the end of each analytical batch). The robustness of the analytical procedure was demonstrated by the tight clustering of QC samples. Relative standard deviation (RSD%) of QC samples was < 30%.

3.2.3 Pre-processing method

Background subtraction, alignment and mass feature detection (mass feature: metabolite with a set retention time, accurate mass and intensity) were all performed in MarkerView software (AB Sciex).

Mass feature detection was obtained based on the following criteria:

1) m/z extraction tolerance: 1 Da;

2) Retention time extraction window: 3-15 min;

3) Retention time tolerance: 0,5 min;

4) Intensity Threshold Feature > $1e^5$;

5) Intensity Noise Threshold: 1000.

The detected mass features were then sequentially subjected to:

1. Normalization based on the internal standard intensity;

2. Normalization based on the sum of the intensity values (the sum of all features of each observation was set to 1).

Pre-processed data were organized in a tidy matrix of dimensions *i* x *j*, where *i* is the number of the instances (corresponding to the number of samples multiplied for three analytical replications) and *j* represents the number of the mass features.

From the analysis of the internal validation sample set (*section 3.2.1*), two different data set were produced, one obtained from the positive mode analysis (POS) and the other from the negative mode (NEG). The first dataset was composed by 207 instances (90 for Italian honey samples; 117 for Eastern European ones) and 2900 features, the second one included the same number of instances, but only 578 features. For the external validation set, the total instances, based on number of the samples described in *section 3.2.1*, were 247, while the number of the features were 2900 and 578, for the positive and negative ionization data sets, respectively.

3.3 Classification approaches

The pre-processed data (*section 3.2.3*) were analysed from the statistical point of view after a first step of feature selection. Feature selection has been performed through a correlation-based algorithm, using two different search methods: (1) a sequential forward-backward selection on the original feature space, resulting in the selection of 23 analytical attributes out of 2900 for POS dataset and 14 out of 578 for the NEG one; and (2) a genetic algorithm-based

search, resulting in the selection of 839 features and 134 for POS and NEG datasets, respectively.

The complete and the reduced datasets have been analysed with the following mathematical classifiers:

- Bayes Net BN with Cooper/Herskovits algorithm with automatic selection of the best number of parents per node;
- K Nearest Neighbor KNN with automatic selection of the best value of K, cross-validation and KDTree search strategy;
- Decision tree J48 with automatic selection of the best confidence factor value;
- Multi Layer Perceptron (MLP) with 1 hidden layer, with automatic selection of the best number of hidden units;
- Support Vector Machine SMO (Sequential Minimal Optimization), with: (1) Pearson Kernel and Platt's scaling, or (2) polynomial kernel, for output class probability estimation;

In addition, the dataset has been tested with a Soft Independent Modelling of Class Analogy (SIMCA) for two classes, matched with Partial Least Squares - Discriminant Analysis (PLS-DA); in this case, prior to the analysis, the datasets have been subjected to the Pareto scaling.

3.4 Internal validation

The internal validation procedure aims at obtaining a classifier (or a set of classifiers) which sub-set of validation parameters, chosen by the experts, matches or overcomes the acceptability thresholds that have been set as minimum performance requirements. To assess the geographic origin of honey, the internal validation has been performed using a binary classification approach (Italian honey vs East Europe), applying a 10-fold cross validation to each combination of the classifiers described in Section 3.3 and the model samples in each of the datasets described in Section 3.2.3 (both POS and NEG), considering each of them as rough data without feature selection, or with feature selection through linear search, or with feature selection using the genetic algorithm (as described in *section 3.3*). After analysing the results of each classifier, trained with each of these datasets, we obtained the best classification results for data from positive ionization mode analysis (POS), followed by correlation-based feature selection (no genetic algorithm) and sequential selection on the original feature space. The following results are, thus, referred to the POS dataset reduced to 207x23 (instances x features) dimensions.
Concerning the thresholds set for the acceptability of the internal validation classification models, we concentrated, on one hand, on the general accuracy, setting a threshold of more than 95% for this parameter and, on the other hand, on parameters to handle the problem of class imbalance affecting the datasets used for this study. For this purpose, we set sensitivity and specificity ratio higher than 0.95 for each class.

The classification parameters for each classifier, providing the best classification results with the selected dataset are summarized as follows:

- Bayes Net BN: K2 algorithm and max. of 2 parents per node;
- K Nearest Neighbor: K=1 and Euclidean distance;
- Decision tree J48: C4.5 Decision Tree with confidence factor of 0.23;
- Multi-layer Perceptron: no. of hidden units = (attribs + classes) / 2;
- Support vector machine: polynomial kernel;
- Soft Independent Modelling of Class Analogy: 4 components for both classes; Partial Least Squares - Discriminant Analysis: 5 component

These classifiers produced the validation results listed in the following tables:

BN:

```
Accuracy             205               99.0338 %
Kappa statistic                          0.9804
Total Number of Instances              207

=== Detailed Accuracy By Class ===

Sensitivity  Specificity  Precision  Recall   F-Measure  MCC      ROC Area   MR       Class
1,000        0,983        0,978      1,000    0,989      0,981    1,000      0,00%    Italian
0,983        1,000        1,000      0,983    0,991      0,981    1,000      0,85%    East Europe
0,990        0,993        0,991      0,990    0,990      0,981    1,000      0,85%    Overall

=== Confusion Matrix ===

   a   b   <-- classified as
  90   0 |   a = Italian
   2 115 |   b = East Europe

=== Mean and variance of classification/misclassification probabilities ===

 Mean:                                                Variance:
     a     b   <-- classified as                          a     b   <-- classified as
   0,995 0,000 |   a = Italian                          0,029 0,000 |   a = Italian
   0,620 0,999 |   b = East Europe                      0,152 0,006 |   b = East Europe
```

KNN:

```
Accuracy              207              100      %
Kappa statistic                          1
Total Number of Instances              207

=== Detailed Accuracy By Class ===

Sensitivity  Specificity  Precision  Recall   F-Measure  MCC      ROC Area   MR       Class
1,000        1,000        1,000      1,000    1,000      1,000    1,000      0,00%    Italian
1,000        1,000        1,000      1,000    1,000      1,000    1,000      0,00%    East Europe
1,000        1,000        1,000      1,000    1,000      1,000    1,000      0,00%    Overall

=== Confusion Matrix ===

   a   b   <-- classified as
  90   0 |   a = Italian
   0 117 |   b = East Europe

=== Mean and variance of classification/misclassification probabilities ===

 Mean:                                                Variance:
     a     b   <-- classified as                          a     b   <-- classified as
   0,995 0,000 |   a = Italian                          0,000 0,000 |   a = Italian
   0,000 0,995 |   b = East Europe                     0,000 0,000 |   b = East Europe
```

J48:

```
Accuracy              201               97.1014 %
Kappa statistic                          0.9409
Total Number of Instances              207

=== Detailed Accuracy By Class ===

Sensitivity  Specificity  Precision  Recall   F-Measure  MCC      ROC Area   MR       Class
0,956        0,983        0,977      0,956    0,966      0,941    0,987      2,22%    Italian
0,983        0,954        0,966      0,983    0,975      0,941    0,987      0,85%    East Europe
0,971        0,967        0,971      0,971    0,971      0,941    0,987      3,07%    Overall

=== Confusion Matrix ===

   a   b   <-- classified as
  86   4 |   a = Italian
   2 115 |   b = East Europe
```

```
=== Mean and variance of classification/misclassification probabilities ===

 Mean:                                                Variance:
     a     b   <-- classified as                          a     b   <-- classified as
   0,999 0,986 |   a = Italian                          0,012 0,010 |   a = Italian
   0,895 0,991 |   b = East Europe                      0,008 0,027 |   b = East Europe
```

MLP:

```
Accuracy              207              100      %
Kappa statistic                          1
Total Number of Instances              207

=== Detailed Accuracy By Class ===

Sensitivity  Specificity  Precision  Recall   F-Measure  MCC      ROC Area   MR       Class
1,000        1,000        1,000      1,000    1,000      1,000    1,000      0,00%    Italian
1,000        1,000        1,000      1,000    1,000      1,000    1,000      0,00%    East Europe
1,000        1,000        1,000      1,000    1,000      1,000    1,000      0,00%    Overall

=== Confusion Matrix ===

   a   b   <-- classified as
  90   0 |   a = Italian
   0 117 |   b = East Europe

=== Mean and variance of classification/misclassification probabilities ===

 Mean:                                                Variance:
     a     b   <-- classified as                          a     b   <-- classified as
   0,993 0,000 |   a = Italian                          0,020 0,000 |   a = Italian
   0,000 0,994 |   b = East Europe                      0,000 0,016 |   b = East Europe
```

SMO:

```
Accuracy              204               98.5507 %
Kappa statistic                          0.9704
Total Number of Instances              207
```

```
=== Detailed Accuracy By Class ===

Sensitivity  Specificity  Precision  Recall   F-Measure  MCC      ROC Area   MR       Class
0,967        1,000        1,000      0,967    0,983      0,971    0,983      1,67%    Italian
1,000        0,967        0,975      1,000    0,987      0,971    0,983      0,00%    East Europe
0,986        0,981        0,986      0,986    0,985      0,971    0,983      1,67%    Overall

=== Confusion Matrix ===

   a   b   <-- classified as
  87   3 |   a = Italian
   0 117 |   b = East Europe

=== Mean and variance of classification/misclassification probabilities ===

 Mean:                                          Variance:
     a     b   <-- classified as                    a     b   <-- classified as
   1,000 1,000 |   a = Italian                     0,000 0,000 |   a = Italian
   0,000 1,000 |   b = East Europe                  0,000 0,000 |   b = East Europe
```

SIMCA/PLS-DA:

```
Accuracy                                95.1691 %
Kappa statistic                          0.9020
Total Number of Instances               207

=== Detailed Accuracy By Class ===

Sensitivity  Specificity  Precision  Recall   F-Measure  MCC       MR        Class
0,956        0,949        0,935      0,956    0,945      0,902     2,22%     Italian
0,949        0,956        0,965      0,949    0,957      0,902     2,56%     East Europe
0,952        0,953        0,952      0,952    0,952      0,902     4,79%     Overall

=== Confusion Matrix ===

   a   b   <-- classified as
  86   4 |   a = Italian
   6 111 |   b = East Europe
```

All the classifiers produced very good results on the POS database, reduced with correlation-based feature selection. In particular, comparing these results with the validation parameters thresholds we have previously set, we can see that all the classifiers trained on this dataset

matched all the requirements and can be considered as a candidate for further testing through external validation.

In the next section of this deliverable, we tested the validation set analysed in positive mode, considering the same features selected for the model dataset, on each of the classifier models trained for this internal validation step.

3.5 External validation

The second validation step (external validation) was carried out by applying the classifiers trained for the internal validation to the validation dataset described in Section 3.2.3, after selecting by hand only the features selected by the automatic correlation-based feature selection and used to train the classification models in the internal validation step (the dimensions of the final data matrix was of 246 x 23).

The method was thus subjected to the validation process for its original aim (see *section 3.1*); the performances of the method are defined in the following tables.

BN:

```
Accuracy              201               81.7073 %
Kappa statistic                          0.5123
Total Number of Instances              246

=== Detailed Accuracy By Class ===

Sensitivity  Specificity  Precision  Recall   F-Measure  MCC      ROC Area    MR      Class
0,650        0,129        0,619      0,650    0,634      0,513    0,859       8,54%   Italian
0,871        0,350        0,885      0,871    0,878      0,513    0,859       9,76%   East Europe
0,817        0,296        0,820      0,817    0,819      0,513    0,859       18,30%  Overall

=== Confusion Matrix ===

   a   b   <-- classified as
  39  21 |   a = Italian
  24 162 |   b = East Europe

=== Mean and variance of classification/misclassification probabilities ===

 Mean:                                          Variance:
     a     b   <-- classified as                    a     b   <-- classified as
   0,971 0,977 |   a = Italian                     0,106 0,078 |   a = Italian
   0,996 0,999 |   b = East Europe                  0,011 0,009 |   b = East Europe
```

KNN:

```
Accuracy             197               80.0813 %
Kappa statistic                          0.463
Total Number of Instances              246

=== Detailed Accuracy By Class ===

Sensitivity  Specificity  Precision  Recall   F-Measure  MCC      ROC Area    MR      Class
0,600        0,134        0,590      0,600    0,595      0,463    0,733        9,76%   Italian
0,866        0,400        0,870      0,866    0,868      0,463    0,733       10,16%   East Europe
0,801        0,335        0,802      0,801    0,801      0,463    0,733       19,92%   Overall

=== Confusion Matrix ===

   a   b   <-- classified as
  36  24 |   a = Italian
  25 161 |   b = East Europe

=== Mean and variance of classification/misclassification probabilities ===

 Mean:                                          Variance:
     a     b   <-- classified as                    a     b   <-- classified as
   0,995 0,995 |   a = Italian                     0,000 0,000 |   a = Italian
   0,995 0,995 |   b = East Europe                  0,000 0,000 |   b = East Europe
```

J48:

```
Accuracy              191               77.6423 %
Kappa statistic                          0.4688
Total Number of Instances              246

=== Detailed Accuracy By Class ===

Sensitivity  Specificity  Precision  Recall   F-Measure  MCC      ROC Area    MR      Class
0,750        0,215        0,529      0,750    0,621      0,483    0,767        6,10%   Italian
0,785        0,250        0,907      0,785    0,841      0,483    0,767       16,26%   East Europe
0,776        0,241        0,815      0,776    0,788      0,483    0,767       22,36%   Overall

=== Confusion Matrix ===

   a   b   <-- classified as
  45  15 |   a = Italian
  40 146 |   b = East Europe
```

```
=== Mean and variance of classification/misclassification probabilities ===

 Mean:                                                Variance:
     a     b   <-- classified as                          a     b   <-- classified as
   1,000 1,000 |   a = Italian                          0,000 0,000 |   a = Italian
   1,000 1,000 |   b = East Europe                      0,000 0,000 |   b = East Europe
```

MLP:

```
Accuracy              203               82.5203 %
Kappa statistic                          0.5676
Total Number of Instances              246

=== Detailed Accuracy By Class ===

Sensitivity  Specificity  Precision  Recall   F-Measure  MCC      ROC Area    MR      Class
0,783        0,161        0,610      0,783    0,686      0,576    0,892        5,28%   Italian
0,839        0,217        0,923      0,839    0,879      0,576    0,892       12,20%   East Europe
0,825        0,203        0,847      0,825    0,832      0,576    0,892       17,48%   Overall

=== Confusion Matrix ===

   a   b   <-- classified as
  47  13 |   a = Italian
  30 156 |   b = East Europe

=== Mean and variance of classification/misclassification probabilities ===

 Mean:                                                Variance:
     a     b   <-- classified as                          a     b   <-- classified as
   0,994 0,832 |   a = Italian                          0,029 0,184 |   a = Italian
   0,982 0,993 |   b = East Europe                      0,029 0,025 |   b = East Europe
```

SMO:

```
Accuracy              208               84.5528 %
Kappa statistic                          0.6078
Total Number of Instances              246
```

```
=== Detailed Accuracy By Class ===

Sensitivity  Specificity  Precision  Recall   F-Measure  MCC      ROC Area    MR      Class
0,783        0,134        0,653      0,783    0,712      0,612    0,824       5,28%   Italian
0,866        0,217        0,925      0,866    0,894      0,612    0,824      10,16%   East Europe
0,846        0,197        0,859      0,846    0,850      0,612    0,824      15,44%   Overall

=== Confusion Matrix ===

   a   b   <-- classified as
  47  13 |   a = Italian
  25 161 |   b = East Europe

=== Mean and variance of classification/misclassification probabilities ===

 Mean:                                          Variance:
     a     b   <-- classified as                    a     b   <-- classified as
   1,000 1,000 |   a = Italian                     0,000 0,000 |   a = Italian
   1,000 1,000 |   b = East Europe                  0,000 0,000 |   b = East Europe
```

SIMCA/PLS-DA:

```
Accuracy                                85.3659 %
Kappa statistic                          0.6032
Total Number of Instances              246

=== Detailed Accuracy By Class ===

Sensitivity  Specificity  Precision  Recall   F-Measure  MCC       MR        Class
0,700        0,903        0,700      0,700    0,700      0,603    15,00%     Italian
0,903        0,700        0,903      0,903    0,903      0,603     4,84%     East Europe
0,815        0,788        0,815      0,815    0,815      0,603    19,84%     Overall

=== Confusion Matrix ===

   a   b   <-- classified as
  42  18 |   a = Italian
  18 168 |   b = East Europe
```

The results of the external validation reveals that the classifiers trained for the internal validation cannot recognize the external validation samples with the same performances; however, except for BN, KNN and SIMCA/PLS-DA classifiers, at least the 75% accuracy value and, for each honey class (Italy and East Europe), a sensitivity of at least 0.75 were reached,

satisfying the thresholds initially requested for these validation parameters. Consequently, the method can be considered promising for future implementations (considering other variables as seasonality, raw materials, manufacturing process, packaging, transport, storage and so on).

From the global evaluation of the results, we have to additionally observe that sensitivity and specificity highlight differences in the recognition of the Italian samples, with respect to the East Europe ones. For example, the BN has a sensitivity of 0,650 for the Italian class and of 0,871 for East Europe, while specificity has values of 0,129 and 0,350 for the same classes, respectively. This situation applies, with more or less impact, also to all the other classifiers, meaning that, in general, the classifiers struggle to identify the origin of the Italian honey samples, while the situation improves when they have to deal with the East European ones. The reasons for such a behaviour should be investigated, together with awareness and attention to the representativeness of sampling planned from the beginning in relation to the purpose and the variables involved, which, at a first glance, could be the main problem leading to the great difference in performance between the internal and the external validation.

Furthermore, even if in the present example we have not defined specific thresholds for all these parameters, this possibility should be carefully considered if the identification (or misclassification) of one class in respect to the other(s) can have a different impact on the basis of the initial purpose of the method, specifically in the case of fraud detection as in the case of the protection of origin, the detection of adulterants and so on. In these cases, it should be recommended to identify also specific threshold for other parameters, for example Specificity, Precision, ROC Area and Misclassification Risk.

4. Achievement of the Deliverable

In the present Deliverable we presented a practical example of the processes required for the in-house validation of a non-targeted method; the validation process was organized in a first step performed on a set of “known” representative samples (internal validation, all the samples were used to train the classifiers) and on a second step (external validation), performed after having collected and analysed another set of “blind” samples (samples unknown to the classifiers, but whose origin is known to the analyst), on which the prediction ability of the developed method was further tested. Through this example we aimed to identify fundamental steps of the process, criticisms and, in a general way, the best practices to be suggested in the next Deliverable of our WP, namely the Deliverable 18.5 “Good practices and methodological guidelines for the validation and application of the untargeted analysis for food authenticity and traceability”.

The case study here considered was selected from our previous activity described in the Deliverable 18.3, and is referred to identification of the geographic origin of Acacia honey samples (Italy *vs* East Europe) through LC/LRMS analytical method coupled with different chemometric classification approaches.

Firstly, we have selected the main validation parameters and measures for assessing the overall quality of the classifier models, in terms of their ability to recognize the correct class labels assigned to the analysed food samples (*section 2. Description of the validation parameters*). Considering that untargeted methods are mainly used to solve classification problems related to food authenticity issues (identification of origin, intended in the widest meaning of the term, adulteration), we have drawn from specific literature the most used parameters able to measure the performances of classification approaches, which were also adaptable to different chemometric techniques (regression models, algorithms, neural networks, …). Particularly, we considered that the evaluation of the classification performance exclusively based on the prediction ability (in term of accuracy or non-error rate) is not satisfying and could even be misleading. Frequently, we observed the difficulty to obtain well balanced classes of samples; in these cases, it would be necessary to consider validation parameters able to manage the unbalanced distribution of the samples, assessing how well the classifier can recognize the positive samples, and how well it can recognize the negative ones. Sensitivity and specificity per class, beside the area under the ROC curve, are parameters useful for overcoming these limitations and more appropriately measuring the performances of the classification model. In addition, in some cases there is the necessity to individuate different weight for the errors, not only because the datasets are often unbalanced, but also because different types of error (e.g., false positive *vs* false negative) might have a different impact on the final outcome, depending on the specific purpose of the study. In these cases, cost-based validation measures should be considered. The costs (or weights) can be assigned on the basis of the balance of each class, or based on the specific knowledge of the application field (assigning a higher cost to misclassification of the negative samples, for example), or combining both of these aspects. As final aspect, because of the probabilistic nature of the classification approaches (a classifier assigns a class to each unknown sample based on a probability value), we suggest to include, among the other validation parameters, the probability assigned to the samples to belong to a specific class; this parameter is a measure of the confidence through which the correct classifications and/or misclassifications are performed. However, due to the high number of samples that we expect to be tested, it could be unpractical to report the probability assigned to each and every sample; thus, in order to overcome this limitation, we suggest at least to calculate and report the mean and the variance of the probabilities determined for each sample typology (true positive, true negative, false positive and false negative).

After having defined the most appropriated validation parameters, the main question to be answered is “what values are accepted to consider the method validated?” Acceptance values for each parameter (or at least for those parameters considered as the most critical ones based on the specific aims of the study designers) should be clearly defined and declared *a priori*, taking into account the purpose of the study and the level of uncertainty eventually admitted. However, it is difficult to generalize and indicate specific thresholds to reach. The level of acceptance should be initially set by the analysts based on the existing knowledge or

on their experience, and depending on the specific purpose of the study. For example, if the validated method needs to be used to classify unknown samples (designation of the origin, identification of adulterated samples,…), the predictive performance should be higher than if the method is used just for preliminary screening (for example in the case of identification of "suspected samples" to be further analysed via targeted methods for confirmations). Furthermore, it can be suggested that for the internal validation (generally performed through the cross-validation technique) the selected parameters should have more restrictive boundaries respect to those set for the external validation step (performed on blind samples, generally displaying more variance than the one present within the internal validation set). Concerning the specific example here presented, the purpose of the study was to develop a preliminary non targeted LC-LRMS method able to discriminate the origin of Acacia honey, but not allowing the identification of unknown samples. For this example, we have set Accuracy and Sensitivity values at least at 95% in order to accept the internal validation, and values >75% to accept the external validation. As a general rule, we can additionally suggest that methods aimed at recognizing unknown samples should present Accuracy levels of at least 95% even after the external validation phase.

Another fundamental aspect that should be carefully considered is how to report the validation performances. A validation report should be always accompanied by the method specifications, including its specific purpose, the main characteristics of the samples and the sampling procedures, the detail of the analytical method, the description of the pre-processing procedures and of the resultant dataset(s), the detailed description of the classification approaches. All these aspects define the application limits of the method, and the validation parameters, when the acceptance levels are reached, are verified within these limits of acceptability.

Finally, as a conclusive consideration, we have to highlight that in the present deliverable we have proposed an example restricted to an *in house* (intra-laboratory) validation protocol. Surely, a more complete validation methodological approach should be implemented including an inter-laboratory phase. Most analytical techniques, and specifically LC/LRMS used in this specific example, present intrinsic limitations that do not allow obtaining equivalent datasets that can be integrated in or evaluated by a unique developed method. Specific trials, requiring more expensive resources (in terms of time, personnel, samples) should be specifically planned in order to overcome these limitations. However, even if currently not applicable in a general way to all the analytical techniques, a specific protocol for inter-laboratory comparisons based on non-targeted NMR analysis will be proposed as annex in the Deliverable 18.5. Based on this protocol, useful to validate unbiased and multi-user NMR based classification tools, other studies can be planned in the future, in order to identify the best operative procedures for different analytical techniques.