

USING MULTIPLE REGRESSION, NEURAL NETWORKS AND SUPPORT VECTOR MACHINES TO PREDICT LAMB CARCASSES COMPOSITION

Filipe Silva^a, Paulo Cortez^a and Vasco Cadavez^b

^a Dep. Information Systems/Algoritmi Centre, University of Minho, 4800-058 Guimarães, PORTUGAL

^b Mountain Research Centre (CIMO), ESA - Polytechnic Institute of Bragança, Apartado 1172, 5301-855 Bragança, PORTUGAL.

Email: filipejgsilva@gmail.com, pcortez@dsi.uminho.pt, vcadavez@ipb.pt

KEYWORDS

Carcass, Multiple Regression, Neural Networks, Support Vector Machines, Tissue.

Abstract

The objective of this work was to use a Data Mining (DM) approach to predict, using as predictors the carcass measurements taken at slaughter line, the composition of lamb carcasses. One hundred and twenty five lambs of Churra Galega Bragançana breed were slaughtered. During carcasses quartering, a caliper was used to perform subcutaneous fat measurements, over the maximum depth of longissimus muscle (LM), between the 12th and 13th ribs (C12), and between the 1st and 2nd lumbar vertebrae (C1). The Muscle (MP), Bone (BP), Subcutaneous Fat (SFP), Inter-Muscular Fat (IFP), and Kidney Knob and Channel Fat (KKCF) proportions of lamb carcasses were computed. We used the *rminer* R library and compared three regression techniques: Multiple Regression (MR), Neural Networks (NN) and Support Vector Machines (SVM). The SVM model provided the lowest relative absolute error for the prediction of BP, SFP and KKCF, while MR presented the best predictions for MP and IFP. Also, a sensitivity analysis procedure revealed the C12 measurement as the most relevant predictor for all five carcass tissues.

INTRODUCTION

The development of a low-cost and expeditious method to predict carcass composition will have applicability for carcasses classification at slaughter line (Cadavez et al., 1999), and for prices definition along the commercialization chain (Cadavez et al., 2002). Carcasses with an optimum composition must have a maximum of lean meat proportion, and optimum organoleptic properties. In this case, the carcass should have a maximum price and if the carcass composition deviates from that optimum its price should be penalized.

Traditionally, the producers estimate the lamb's carcass composition through subjective, thus imprecise, methods such as visual assessment and palpation. However, at slaughter line the methodology to predict the car-

casses composition should be accurate, fast, and automated. Data Mining (DM) techniques aim at extracting high-level knowledge from raw data (Witten and Frank, 2005) and can represent an interesting alternative for predicting carcass composition, which can be set by collecting several carcasses parameters at slaughter line.

Typically, these parameters are collected during the slaughtering process or within the first 24 hours after slaughtering. Indeed, several studies have adopted such data-driven approach based on Multiple Regression (MR) models (Hopkins, 2008; Cadavez, 2009), using as independent (or input) variables the carcass weight, in combination with subcutaneous fat depth (Hopkins et al., 2008), *longissimus* muscle depth, and total tissue thickness (Kirton et al., 1984; Hopkins et al., 2008). Yet, these linear models may fail when nonlinear relationships are present in the data and when predictors suffers of multiple collinearity (Cadavez, 2009). In such scenarios, there is a need for alternative modeling techniques, such as the more flexible Neural Networks (NN) and Support Vector Machines (SVM) (Hastie et al., 2008).

In this work, we follow a DM approach to predict the composition of lamb carcasses based on noninvasive carcass measurements that are easy to collect after slaughtering. In particular, we compare three regression models (MR, NN and SVM).

MATERIALS AND METHODS

Lamb Carcass Data

One hundred and twenty five lambs of Churra Galega Bragançana (42 females, and 83 males), randomly selected from the experimental flock of the Escola Superior Agrária de Bragança, were used. Lambs were slaughtered after 24 hour fast in the experimental slaughterhouse at the Escola Superior Agrária de Bragança, and carcasses were weighted approximately 30 minutes after slaughter in order to obtain the Hot Carcass Weight (HCW). Carcasses were halved through the center of the vertebral column, and the Kidney Knob and Channel Fat (KKCF) was removed and weighed. During quartering, tissue measurements were performed with a caliper on maximum LM depth (mm) and subcuta-

Table 1: Dataset main attributes

| Attribute | Description | Domain |
|-------------|--------------------------------------------------------------------------|--------------|
| sex | Lamb sex | $\{1, 2\}^a$ |
| HCW | Hot Carcass Weight (kg) | [5.3, 23.3] |
| C1 | Subcutaneous fat at 1 st lumbar vertebrae (mm) | [0.4, 5.9] |
| C12 | Subcutaneous fat at 12 th rib (mm) | [0.5, 7.1] |
| B1 | <i>Longissimus</i> muscle depth at 1 st lumbar vertebrae (mm) | [14.9, 37.7] |
| B12 | <i>Longissimus</i> muscle depth at 12 th rib (mm) | [13.6, 33.6] |
| MP | Muscle proportion (mass fraction) | [0.47, 0.68] |
| BP | Bone proportion (mass fraction) | [0.14, 0.26] |
| SFP | Subcutaneous fat proportion (mass fraction) | [0.02, 0.16] |
| IFP | Intermuscular fat proportion (mass fraction) | [0.06, 0.16] |
| KKCF | Kidney knob and channel fat proportion (mass fraction) | [0.01, 0.11] |

^a 1 - Male, 2 - Female

neous fat thickness (mm) between the 12th and 13th ribs (B12 and C12, respectively), 1st and 2nd lumbar vertebrae (B1 and C1, respectively). The dataset main attributes are shown in Table 1.

Each carcass was then dissected into muscle, subcutaneous fat, intermuscular fat, bone, and remainder (major blood vessels, ligaments, tendons, and thick connective tissue sheets associated with muscles), and the Muscle (MP), Bone (BP), Subcutaneous Fat (SFP), Intermuscular Fat (IFP), and Kidney Knob and Channel Fat (KKCF) proportions of lamb carcasses were computed.

Models Evaluation

A regression dataset D is made up of $k \in \{1, \dots, N\}$ examples. Each example maps an input vector (x_1^k, \dots, x_I^k) to a given target y_k . The error for a given k is: $e_k = y_k - \hat{y}_k$, where \hat{y}_k represents the predicted value for k input pattern. The regression models performance was evaluated using the *Relative Absolute Error (RAE)* and coefficient of determination (R^2) (Witten and Frank, 2005):

$$RAE = 1/N \times \sum_{i=1}^N \frac{|y_i - \hat{y}_i|}{|y_i - \bar{y}_i|} \quad (1)$$

$$R^2 = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y}_i)^2} \quad (2)$$

where \bar{y} is the mean of the output variable. The RAE statistic is scale independent, and values close to 100% corresponds to a model that has a similar performance as the naive average predictor (i.e. $\hat{y}_i = \bar{y}_i$). The lower the RAE, the better is the regression model, thus the ideal regression model presents a value close to 0%. A high the R^2 value suggests that the predictions and target values are highly correlated; i.e. a high proportion of the variability of the target is followed by the predictions. The ideal model presents an $R^2 = 1$.

To estimate the generalization capability of the regression models, a 10-fold cross-validation procedure was used. With this procedure one subset is tested each time and the remaining data are used for fitting the model. The process is repeated sequentially until all subsets have been tested. Therefore, under this scheme, all data are used for training and testing. Since the results can depend on the random split used to set the 10 folds, we also apply 20 runs to each 10-fold process, in a total of $20 \times 10 = 200$ experiments for each tested configuration. Statistical confidence will be given by the t-student test at the 95% confidence level.

Learning Models

A Multiple Regression (MR) model is defined by the equation (Hastie et al., 2008):

$$Y_i = \beta_0 + \sum_{i=1}^n \beta_i X_i, i = 1, 2, \dots, n \quad (3)$$

where: Y_i is the response (carcass tissue proportion) in the i th case, X_i is the value of the independent variable in the i th case and β_i are the regression coefficients. This model is easy to interpret and has been widely used for the prediction of carcasses composition.

Neural Networks (NNs) are connectionist models that found their inspiration on the behavior of the human brain. In particular, the multilayer perceptron is the most popular NN architecture, and can be defined as a feedforward network where processing neurons are grouped into layers and connected by weighted links (Haykin, 1999). This study considered the multilayer perceptrons with one hidden layer of H hidden nodes, with logistic activation functions and one output node with a linear function (Hastie et al., 2008). Since the NN cost function is nonconvex (with multiple minima), $NR=3$ trainings was applied to each neural configuration, and the NN with the lowest fitted error was selected. Under this setting, the NN performance depends

on the value of H . If $H = 0$, the model is equivalent to the MR. When increasing H , a more complex mapping is performed, yet an excess value of H will overfit the data, leading to generalization loss.

Support Vector Machines (SVM) present theoretical advantages over NN, such as the absence of local minima in the model optimization phase. In SVM regression, by using a nonlinear mapping, the input $x \in \mathbb{R}^I$ is transformed into a high m -dimensional feature space. Then, the SVM finds the best linear separating hyperplane in the feature space. In this work, the nonlinear transformation is achieved by adopting the popular gaussian kernel, which presents less parameters than other kernels (e.g. polynomial): $K(x, x') = \exp(-\gamma\|x - x'\|^2)$, $\gamma > 0$. Also, we adopted the commonly used ϵ -insensitive loss function, which sets an insensitive tube around the residuals and the tiny errors within the tube are discarded. Under this setup, the SVM performance is affected by three parameters: γ , ϵ and C (a trade-off between fitting the errors and the flatness of the mapping). To reduce the search space, the first two values will be set using the heuristics (Cherkassy and Ma, 2004): $C = 3$ (for a standardized output) and $\epsilon = \hat{\sigma}/\sqrt{N}$, where $\hat{\sigma} = 1.5/N \times \sum_{i=1}^N (y_i - \hat{y}_i)^2$ and \hat{y}_i is the value predicted by a 3-nearest neighbor algorithm. The kernel parameter (γ) produces the highest impact in the SVM performance, with values that are too large or too small leading to poor predictions.

To adjust the NN and SVM hyperparameters (e.g. H and γ) a grid search (with $H \in 1, 2, \dots, 8$ and $\gamma = 2^{-13}, 2^{-11}, \dots, 2^1$, in a total of 8 searches per model) was used. An internal 3-fold (using only training data) was used to select the best hyperparameter. Then, the best model was retrained with all training data (as defined by the external 10-fold validation scheme).

The relative importance of the predictors (or inputs) for a given DM model can be estimated by using a sensitivity analysis procedure (Cortez et al., 2009). This procedure measures how the responses are affected when all inputs are hold at their average values except x_a , which varies through its entire range. The attribute x_a is considered more relevant if it produces a higher variance in the responses. A more detailed input influence analysis is given by the Variable Effect Characteristic (VEC) curve, which plots the x_a values (x -axis) versus the \hat{y}_a responses (y -axis).

RESULTS AND DISCUSSION

All experiments reported in this study were conducted using the **rminer** library, which facilitates the application of DM techniques in the **R** simulation tool Cortez (2010). The test set results are shown in Tables 2 and 3 in terms of the mean RAE (equation 1) and R^2 (equation 2) values and respective 95% t-student confidence intervals. In the tables, the best values are in **bold**,

while underline denotes a statistical significance (P<0.05) under a pairwise comparison against other methods.

The MR model presented the lowest RAE for MP (RAE=59.4%, P<0.05) and for IFP (RAE=64.1%, P<0.05). The SVM model presented the lowest RAE for BP (46.1%, P<0.05), for KKCF (51.5%, P<0.05), and for SFP (42.2%, P<0.05). However, it is important to notice that for SFP and IFP prediction the differences among regression models were not statistically significant. The RAE results show an overall improvement of around 36% (IFP) to 58% (SFP) when compared with the naive average predictor. The SVM modeling gives the best predictions for BP, SFP and KKCF, while MR achieves the lower RAE when predicting the MP and the IFP. The NN model only outperforms the MR model for prediction of BP and SFP.

Table 2: RAE values (in %) for predicting lamb carcass composition (test set results)

| | MR | NN | SVM |
|-------------|-------------------|------------|-------------------|
| MP | 59.4 ± 0.3 | 63.0 ± 1.6 | 60.1 ± 0.7 |
| BP | 48.4 ± 0.3 | 47.0 ± 0.4 | 46.1 ± 0.3 |
| SFP | 43.1 ± 0.3 | 42.5 ± 0.6 | 42.2 ± 0.4 |
| IFP | 64.1 ± 0.3 | 64.5 ± 0.8 | 65.8 ± 0.8 |
| KKCF | 53.1 ± 0.5 | 57.7 ± 2.0 | 51.5 ± 0.5 |

The R^2 values are in general higher for the models that present the lowest RAE values (e.g. $R^2 = 0.71$, P<0.05, for KKCF and SVM) and range from around 0.6 to 0.8 for the best regression models. These R^2 values are higher than that reported by Cadavez (2009) for prediction of MP ($R^2=0.42$) in a study with male lambs of Suffolk and Churro Galego Bragançano breeds. Similarly, Hopkins et al. (2004) reported R^2 > 60% for models predicting the carcasses lean meat proportion, in a study with lambs of several breeds, where video image analysis was also tested.

Table 3: R^2 values for predicting lamb carcass composition (test set results)

| | MR | NN | SVM |
|-------------|--------------------|--------------------|--------------------|
| MP | 0.65 ± 0.00 | 0.55 ± 0.11 | 0.63 ± 0.01 |
| BP | 0.75 ± 0.00 | 0.78 ± 0.00 | 0.78 ± 0.00 |
| SFP | 0.79 ± 0.00 | 0.80 ± 0.02 | 0.80 ± 0.00 |
| IFP | 0.58 ± 0.00 | 0.58 ± 0.01 | 0.57 ± 0.01 |
| KKCF | 0.70 ± 0.01 | 0.54 ± 0.09 | 0.71 ± 0.00 |

To show the quality of the results achieved, Figure 1 plots the observed vs predicted scatter plots for the best regression models. In the plots, the majority of the plots are close to the diagonal line, which denotes a perfect forecast. However, in some cases there are larger errors when predicting the extreme values. For instance, the KKCF predictions underestimate the target values that are close to the maximum KKCF value. Also, a trend to

the overestimation of the lowest KKCF, SFP and IFP values can be observed (Figure 1).

The relative importance (as defined by the sensitivity analysis) of the predictors of carcass composition for best models is presented in Figure 2. The C12 measurement is the most important predictor for all tasks, with an influence that ranges from around 25% (BP and MR model) to 80% (MP and MR model). These results are in accordance with those attained by Cadavez (2009), where fat measurements dominate the models for prediction of MP.

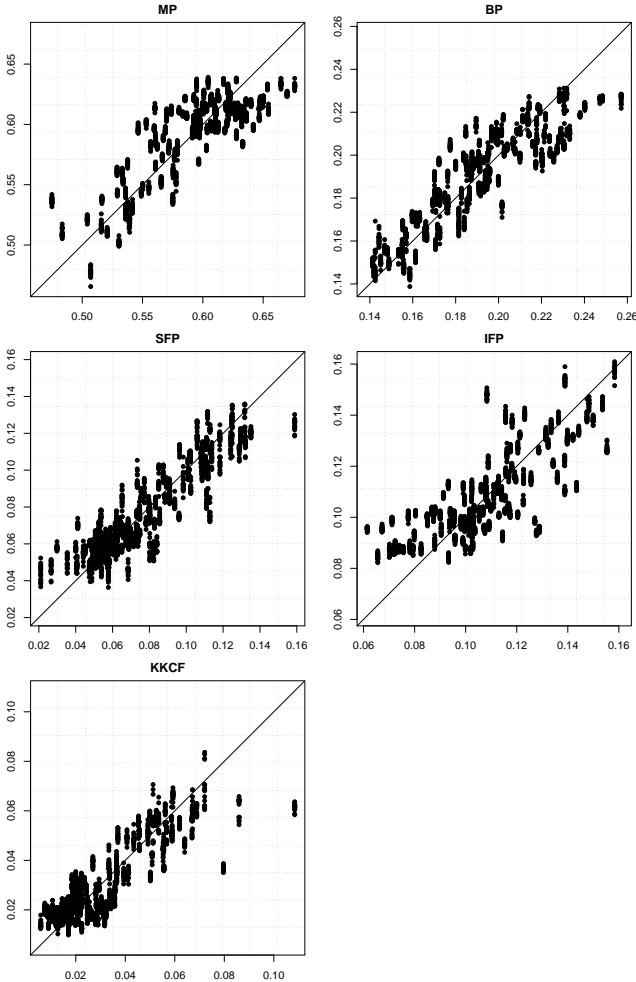


Figure 1: Scatter plots for the best regression models (x -axis – target values, y -axis – predictions)

The relative importance of the remaining input variables varied from task to task. For example, the sex was the second most relevant predictor for the KKCF prediction model (SVM), while it was the least important predictor for the SFP prediction model. Figure 3 plots the VEC curves for the C12 input and MP and SFP models. In the former graph, there is a negative linear influence of C12. In other words, the increase in the C12 measurement leads to a decrease in the carcasses MP. Regarding the latter VEC, the influence of C12 in the overall SFP

is positive. In this case, the SVM measured a nonlinear (i.e. parabola shape) influence.

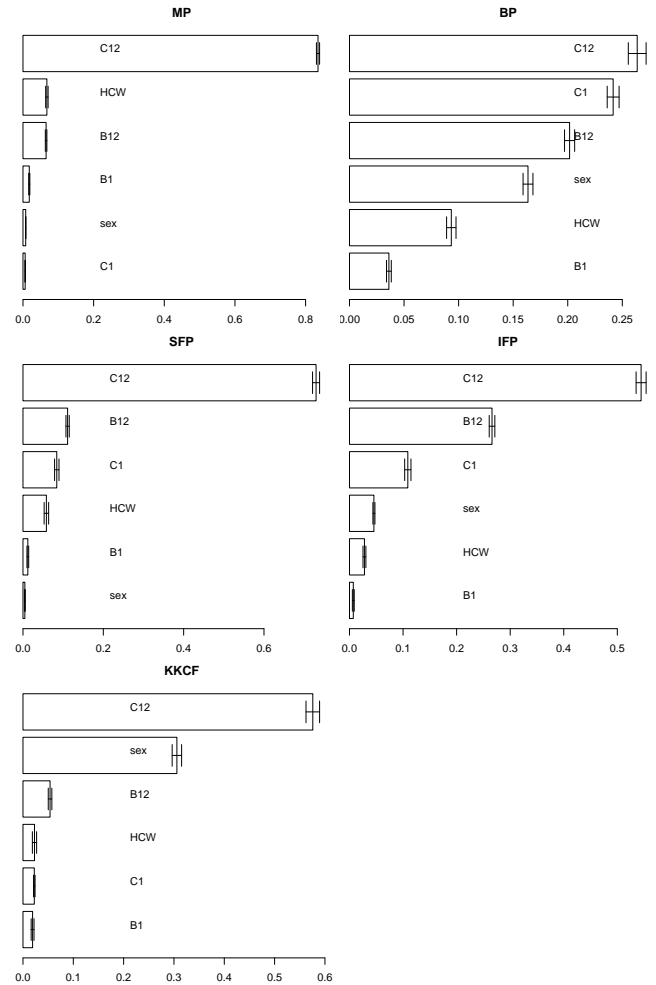


Figure 2: Relative importance of the predictors.

CONCLUSIONS

The SVM modeling technique provides the lowest RAE values for BP, SFP and KKCF prediction, while MR gives the best predictions for MP and IFP. The R^2 values of the best predictions range from 0.6 to 0.8. Sensitivity analysis revealed the C12 measurement (subcutaneous fat depth at 12th rib) as the most relevant predictor for the prediction of all five carcass tissue. It is also important to notice that C12 measurement presents a nonlinear and positive influence when modeling the SFP.

These results are relevant to the animal science domain, helping in the understanding of how carcass tissues measurements affect the carcasses composition. In addition, these findings can have an impact in the meat industry, since the models developed to predict the carcass composition can be integrated into decision support systems,

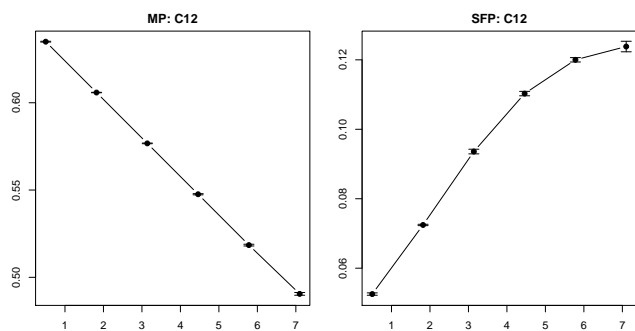


Figure 3: VEC curve showing the C12 influence on MP (left) and SFP (right) prediction models.

in order to implement a system capable of defining carcass prices based on the prediction of carcasses tissues proportion by objective models. In future research, we will also design methods to select regression models that provide better predictions at the extreme values, since such models would provide a more accurate detection of high/low quality meat.

References

- Cadavez V.; Rodrigues S.; Pereira E.; Delfa R.; and Teixeira A., 2002. *Predicción de la composición de la canal de cabritos por ultrasonografía in vivo*. ITEA, 98A(1), 39–50.
- Cadavez V.A.P., 2009. *Prediction of lean meat proportion of lamb carcasses*. *Archiva Zootechnica*, 12, no. 4, 46–58.
- Cadavez V.A.P.; Teixeira A.; and Delfa R., 1999. *Utilización de ultrasonidos junto con el peso vivo y el peso de la canal caliente para la estimación del peso de las piezas de carnicería en corderos de raza churra Galega Bragançana: Comparación de sondas de 5 y 7,5 MHz*. *Producción Ovina y Caprina*, XXIV, 425–432.
- Cherkassy V. and Ma Y., 2004. *Practical Selection of SVM Parameters and Noise Estimation for SVM Regression*. *Neural Networks*, 17, no. 1, 113–126.
- Cortez P., 2010. *Data Mining with Neural Networks and Support Vector Machines using the R/rminer Tool*. In P. Perner (Ed.), *Advances in Data Mining, Proceedings of 10th Industrial Conference on Data Mining*. LNAI 6171, Springer, Berlin, Germany, 572–583.
- Cortez P.; Teixeira J.; Cerdeira A.; Almeida F.; Matos T.; and Reis J., 2009. *Using Data Mining for Wine Quality Assessment*. In J. Gama; V.S. Costa; A.M. Jorge; and P. Brazdil (Eds.), *Discovery Science*. Springer, *Lecture Notes in Computer Science*, vol. 5808. ISBN 978-3-642-04746-6, 66–79.
- Hastie T.; Tibshirani R.; and Friedman J., 2008. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, NY, USA, 2nd ed.
- Haykin S., 1999. *Neural Networks: A Comprehensive Foundation*. Prentice Hall, New Jersey, USA, 2.
- Hopkins D.L., 2008. *An industry applicable model for predicting lean meat yield in lamb carcasses*. *Australian Journal of Experimental Agriculture*, 48, 757–761.
- Hopkins D.L.; Ponnampalam E.N.; and Warner R.D., 2008. *Predicting the composition of lamb carcasses using alternative fat and muscle depth measures*. *Meat Science*, 78, 400–405.
- Hopkins D.L.; Safari E.; Thompson J.M.; and Smith C.R., 2004. *Video image analysis in the Australian meat industry - precision and accuracy of predicting lean meat yield in lamb carcasses*. *Meat Science*, 67, 269–274.
- Kirton A.; Woods E.; and Duganzich D., 1984. *Predicting the fatness of lamb carcasses from carcass wall thickness measured by ruler or by a total depth indicator (TDI) probe*. *Livestock Production Science*, 11, no. 2, 185–194.
- Witten I. and Frank E., 2005. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, San Francisco, CA.

AUTHOR'S BIOGRAPHY

FILIPE SILVA is an M.Sc. student in Information Systems at the University of Minho.

PAULO CORTEZ (Ph.D) is an Assistant Professor at the Dep. of Information Systems and Researcher at the Algoritmi centre of U. Minho. He is co-author of more than sixty publications in international conferences, journals and book chapters. Research interests: Business Intelligence and Data Mining; Neural and Evolutionary Computation; Forecasting (see: <http://www3.dsi.uminho.pt/pcortez>).

VASCO CADAVEZ (Ph.D) is an Adjutant professor in the Dep. of Animal Science at Polytechnic Inst. of Bragança (IPB). His research interests include carcass and meat quality, with special emphasis on the development of models for objective prediction of carcasses composition. Since 2009, he is Principal Investigator of Food Safety and Technology research group at Mountain Research Centre (CIMO).