Cranfield University

O. OZCAN

# DEVELOPMENT OF MACHINE LEARNING REGRESSION MODELS TO ASSESS FRESHNESS IN CHICKEN BREAST FILLETS

SCHOOL OF WATER, ENERGY AND ENVIRONMENT

Applied Bioinformatics

MSc

Academic Year: 2018 – 2019

Supervisors: Dr Fady Mohareb, Dr Maria Anastasiadi

August 2019

# Table of Contents

List of abbreviations ..... 2
Abstract ..... 4
1. Introduction ..... 4
2. Materials and methods ..... 5
*2.1 Sample preparation and experimental design* ..... 5
*2.2 Microbiological analyses* ..... 6
*2.3 Multispectral imaging* ..... 6
*2.4 Fourier transform infrared spectroscopy* ..... 6
*2.5 A closer look at the batches* ..... 7
*2.6 Statistical modelling* ..... 8
*2.7 Performance metrics* ..... 9
3. Results ..... 9
*3.1 Multispectral imaging data results* ..... 9
*3.2 Fourier transform infrared spectrum data results* ..... 10
4. Discussion ..... 11
5. Conclusions ..... 12
Acknowledgements ..... 12
References ..... 12
Appendix A ..... 14
Appendix B ..... 14

# List of abbreviations

| | |
|---|---|
| FTIR | Fourier transform infrared spectrum |
| MSI | Multispectral imaging |
| TVC | Total viable counts |
| CFU | Colony forming unit |
| SSO | Specific spoilage organisms |
| ESO | Ephemeral spoilage organisms |
| R | R language and environment for statistical computing and graphics |
| PCA | Principal component analysis |
| RMSE | Root mean squared error |
| MAE | Mean absolute error |
| $R^2$ | R-squared, coefficient of determination |
| MLP | Multilayer perceptron |
| SVMRFE | Support vector machine recursive feature elimination |
| svmLinear | Support vector machine with linear kernel |
| svmRadial | Support vector machine with radial kernel |
| rf | Random forest |
| knn | K-nearest neighbour |
| pcr | Principal component regression |
| ridge | Ridge regression |
| ANN, nnet | Artificial neural network |
| lars | Least-angle regression |
| pls | Partial-least squares |

# Illustration index

Figure 1 - TVC over storage times, grouped by temperature. ..... 6
Figure 2 - Typical MSI data collected from fresh and spoiled samples ..... 6
Figure 3 – Typical FTIR data collected from fresh and spoiled samples with regions of interest highlighted. ..... 7
Figure 4 - Pre-processed typical FTIR data collected from fresh and spoiled samples. ..... 7
Figure 5 - PCs 1 and 2 plotted for MSI data. ..... 8
Figure 6 - PCs 1 and 2 plotted for FTIR data. ..... 8
Figure 7 - Outline of the statistical modelling procedure. ..... 8
Equation 1 – Equation for root mean squared error. ..... 9
Equation 2 – Equation for mean absolute error. ..... 9
Equation 3 – Equation for R-squared. ..... 9
Equation 4 – Equation for accuracy. ..... 9
Figure 8 - MSI intra-batch results ranked by validation accuracy. ..... 10
Figure 9 - MSI batch-on-batch results ranked by validation accuracy. ..... 10
Figure 10 - FTIR intra-batch results ranked by validation accuracy. ..... 10
Figure 11 - FTIR batch-on-batch results ranked by validation accuracy. ..... 10
Figure 12 - Plot of predicted vs actual for the highest performing models for MSI batch-on-batch testing. ..... 11
Figure 13 - Plot of predicted vs actual for the highest performing models for FTIR batch-on-batch testing. Red lines indicate the boundaries for predictions within 1 $\log_{10}$ cfu. ..... 11
Figure 14 - Demonstration of possible over-fitting for FTIR with Random Forest. ..... 12

# Development of machine learning regression models to assess freshness in chicken breast fillets

Onur Ozcan

onur.ozcan@cranfield.ac.uk

## ABSTRACT

The aim of this work was to evaluate the potential of Fourier transform infrared spectroscopy (FTIR) and multispectral imaging (MSI) as rapid and accurate techniques for monitoring spoilage in chicken breast fillets. FTIR (4000 to 400 $cm^{-1}$) and MSI spectra (405 to 970 nm) analysis was performed on two batches of chicken fillet, which include a total of 215 samples stored at 0°, 5°, 10°, 15° with storage periods ranging between 0 to 518 hours. Moreover, samples were subjected to microbiological analyses to enumerate total microbial counts (TVC), which ranged from 2.69 to 7.77 $log_{10}$ cfu/g. Several machine learning models were trained, calibrated and validated to correlate TVC with the spectral data. The performance evaluation of the models were based on statistical indices (i.e. root mean square error) and prediction accuracy. Models were ranked on their performances on both intra-batch testing where the models are trained on one part of a batch and validated on the remaining, and inter-batch testing where they are trained on one whole batch and validated on the other batch. Artificial neural network (ANN) and least-angle regression (lars) were identified as the models with the highest performances overall. Furthermore, ANN models built on the MSI data achieved the highest accuracy for intra-batch testing with ~80.6% accuracy across all batches. On the other hand, least-angle regression excelled with the FTIR data on inter-batch testing with ~76.2% accuracy. The results suggested that FTIR platform tends to surpass MSI on predicting samples from different batches while MSI is effective for monitoring samples from the same batch.

## 1. Introduction

The demand for high-quality food today is greater than ever (Van Wezemael *et al.*, 2010). Consumers expect their foods to have superior sensorial quality and good image. Although the consumer perception of food quality can be deemed subjective, influenced by cultural values and personal sensory acuity of the individual, food spoilage is regarded as the primary factor whether if the products are preferred over others (Nychas *et al.*, 2008). Among different food industries, meat industry is particularly vulnerable to quality issues since meat products are susceptible to spoil relatively fast with their nutrient rich form which acts as a suitable medium for many pathogens and spoilage microbes to inhabit. Therefore, the meat industry needs swift methods to determine product quality and to establish appropriate processing measures while predicting shelf life (Mohareb *et al.*, 2016).

Meat spoilage can be defined as changes that occur in available substrata during the accumulation of certain microbial associations, namely specific spoilage organisms (SSO) or ephemeral spoilage organisms (ESO) (Nychas, Marshal and Sofos, 2007). The development of these microbiological organisms can start at slaughter and proceed throughout the product's life cycle, namely chilling, transport and storage (Nychas *et al.*, 2008). Even though this build-up and its metabolic by-products may not be harmful to human health, they tend to give products repulsive outlook and smell. The current procedure for spoilage evaluation relies on either laboratory work which involve enumerating bacterial colonies by microbiological methods, or sensory panels which consist of trained personnel who estimate spoilage levels. Microbiological methods tend to be destructive and unreliable, while sensory panels can be considered subjective depending on the individual preferences of the panellists. In addition, both methods fall short in delivering rapid and accurate results while being financially and logistically un-feasible (Daalgard, 1995).

Recently, many machine learning algorithms have been either developed or optimised to make better use of the current computational power, such as support vector machines (SVM), artificial neural network (ANN) and others (Williams, 2003). These algorithms combined with the sample data, which is collected via various analytical platforms, is used to predict spoilage through either classification or regression methods. Classification attempts to assign new samples into categories, such as fresh, semi-

fresh and spoiled. On the other hand, regression attempts to predict a continuous value, usually total or specific bacterial counts.

In the last two decades, new, effective, rapid and non-destructive analytical platforms for collecting sample data have been developed. Among these, multispectral imaging (MSI) and Fourier transform infrared spectra (FTIR) data display great potential. FTIR platform measures the absorbance of infrared light at various wavenumbers, often ranging from 4000 to 400 $cm^{-1}$, which can give insight on sample composition and structure. Further details on FTIR can be found elsewhere (Smith, 2011). However, it is shown that only specific wave bands should be incorporated into quantitative analysis, depending on the purpose of the research (Lasch and Naumann, 2015). The methods for selecting these regions often vary between studies, commonly utilised methods include chemometrics (i.e. PCA), wave-based intensity analysis, and feature selection techniques (i.e. stepwise selection).

In early 2000's, FTIR was successfully utilised to separate biochemical footprints of microbial cells, which led to classification of bacterial species (Rodriguez-Saona *et al.*, 2001). Studies on food quality assessment started soon after with efforts of correlating microbial counts on beef, pork and poultry. One of the earliest studies showed that partial-least-squares regression (PLS-R) was effective at predicting total microbial counts (TVC) on raw chicken breast meat (Ellis and Broadhurst, 2002). The study pointed that the absorbance between 1315 and 1000 $cm^{-1}$ had the most significance. Another study with PLS-R showed similar results on minced beef, however, the wave lengths that were reported as noteworthy were 1714 to 1710 $cm^{-1}$, 1614 to 1211 $cm^{-1}$ and 1031 to 1000 $cm^{-1}$ (Ammor, Argyri and Nychas, 2009). Besides TVC, it is shown that specific bacterial colony counts such as lactic acid bacteria can also be a strong indicator of spoilage. Papadopoulou *et al.*, 2011, on their study on minced pork meat, reported correlations above 0.80 with lactic acid bacteria. In addition, the study pointed out that spectral data between 973-971 $cm^{-1}$, 997-985 $cm^{-1}$, 1041-1016 $cm^{-1}$, 1371-1292 $cm^{-1}$, 1486-1388 $cm^{-1}$, 1540-1529 $cm^{-1}$ and 1726-1697 $cm^{-1}$ were the most important as they corresponded to certain biological structures such as p-o bonds, proteins and so on (Pedersen *et al.*, 2003; Ellis, Broadhurst and Goodacre, 2004). On another study which focused on chicken fillets, Vasconcelos, Saraiva and de Almeida, 2014 utilised FTIR spectral data from 144 samples at different temperatures and ran PLS-DA classification and PLS-R regression on TVC, lactic acid bacteria and *pseudomonas.* The classification had 73.3% accuracy and a reasonable correlation was found between TVC and sensory freshness with root square mean error (RMSE) of 0.789. In addition, it is reported that the spoilage was observable when TVC exceeded 8 log cfu $g^{-1}$**.** Moreover, the study found that on average, samples were considered spoiled after 240 hours on 3 degrees, 168 hours on 8 degrees, 7 hours on 30 degrees storage. Furthermore, it is pointed out that spectral data between 1408-1370 $cm^{-1}$ and 1320-1305 $cm^{-1}$ linked to amides and amines and strongly correlated to spoilage. Artificial neural network (ANN) is another prevalent algorithm which is quickly gaining popularity due to massive computational power improvements in recent years. A study on beef fillets achieved 90.5% classification accuracy using multilayer perceptron (MLP) neural network, utilising FTIR spectra between 1800 and 1000 $cm^{-1}$ (Argyri *et al.*, 2010).

On the other hand, imaging techniques have long been used for visual evaluation of food quality by processing colour, shape, size and surface texture of the samples (Girolami *et al.*, 2013). In fact, MSI is proving to produce excellent results as it incorporates traditional imaging and spectroscopy to retrieve both spatial and spectral information from the samples (Huang *et al.*, 2015). This is done by recording surface reflections collected from wave bands often ranging from 400 to 1000 nm (Dissing *et al.*, 2013). A relatively recent study on pork meat achieved clear separation between samples by utilising 18 wave bands and unsupervised K-means clustering. Same study also applied PLS-R on TVC and reported a root mean square error (RMSE) of 0.551 for TVC values ranging from 5.36 to 9.68 log cfu $g^{-1}$ (ibid.). In contrast, another study focused on three wavebands at 1280, 1440 and 1660 nm. Back propagated adaptive boosting (BP-AdaBoost) is compared to partial least squares regression (PLS-R) using total volatile basic nitrogen content as the prediction value. BP-AdaBoost uses back propagated artificial neural networks (BP-ANN) as the weak learning algorithm and improves its own performance accordingly. The final BP-AdaBoost regression model achieved root mean square error (RMSE) of 6.9439 mg/100 g while PLS-R lacked behind with 8.67 mg/100 mg (Huang *et al.*, 2015).

Although there exists some research on utilising machine learning to predict spoilage of beef, pork, and poultry, specific research on chicken breast fillets utilising MSI and FTIR platforms is limited. This is specifically relevant because chicken consumption has increased in regards to other products in Europe, due to its high protein and low-fat content, along with being low-cost and convenient to include in the modern diet (Magdelaine, Spiess and Valceschini, 2008). Therefore, the objective of this study is focused on comparing the effectiveness of different machine learning models in tandem with FTIR and MSI sample data, to give a comprehensive review on which combination is suitable for detection of spoilage in chicken breast fillets.

## 2. Materials and methods

### *2.1 Sample preparation and experimental design*

The data from both MSI and FTIR platforms, and microbiological analyses were supplied by Laboratory of Microbiology and Biotechnology of Foods, of the Agricultural University of Athens, Greece. Two independent batches of chicken breast fillets, 115 and 100 in sizes respectively, were obtained from a local market and transported to the laboratory. 215 samples were stored at different isothermal conditions (0°, 5°, 10° and 15°C) and were subjected to microbiological analyses, multispectral imaging acquisition (MSI), Fourier transform infrared spectroscopy measurements (FTIR) at certain time intervals ranging from 0 to 518 hours.

For the MSI platform, samples were transferred into Petri dishes and images were acquired using the VideometerLab system (Carstensen and Hansen, 2003; http://www.videometer.com). The data consisted of 18 mean intensity values collected from wavenumbers between 405 and 970 nm and their corresponding standard deviations. As for the FTIR data, spectra was captured over the range of 4000 to 400 $cm^{-1}$ with ~1 $cm^{-1}$ intervals by

utilising a Nicolet 6700 FTIR Spectrometer. Collected FTIR spectra involved 3736 features.

All subsequent analyses and modelling were conducted via R language and environment for statistical computing and graphics (R Foundation for Statistical Computing., 2018) and its relevant packages, which are listed at Appendix A.

### *2.2 Microbiological analyses*

Total bacterial count of samples grouped by temperature over storage times can be seen at *Figure 1*. It can be observed that TVC values are positively correlated with both temperature and storage time. However, the batches deviate from one another in some areas. Although the mean TVC values for both batches are similar (5.079 and 5.130), batch 1 covers a larger range with more deviation (std: 1.568 vs 1.348). There is also a noticeable difference where batch 1 shows higher correlation (Pearson correlation) between TVC and storage time (0.587 vs 0.427), and temperature (0.438 vs 0.387). Details on other descriptive statistics can be found at Appendix B.

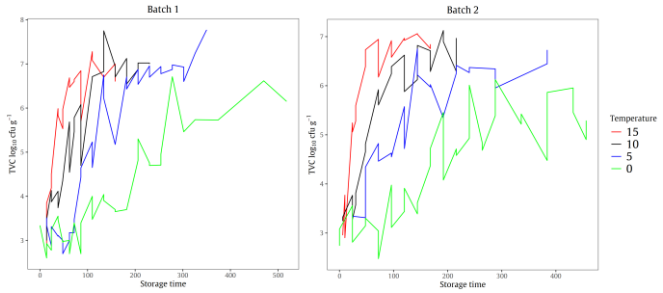



Figure 1 - TVC over storage times, grouped by temperature.

### *2.3 Multispectral imaging*

Typical MSI data collected from fresh and spoiled samples are shown at *Figure 2*. A peak can be seen around 12nd mean (870 nm) which could provide useful information on biochemical changes during the spoilage process.

Prior to statistical modelling, dealing with outliers is a usual preparatory procedure when the goal is to apply machine learning in a scenario where anomalies are not the main influential factor. Outliers in the data could occur due to experimental errors which could severely affect a model's prediction ability. Cook's distance test is a multivariate method which is frequently used to identify outliers. As a general rule of thumb, observations with a Cook's distance above 4 times the mean are considered possible outliers (Stevens, 1984). A visualization of cook's distance test on batch 1 can be seen at Appendix B. Five and seven strong candidate outliers for batch 1 and 2 respectively was observed by utilising this test and PCA was used as a subsequent analysis for confirmation. Several weak possible outliers were also observed, but no action was carried out to capture more variance since recursively removing weak outliers might limit the models' generalisation ability, which could lead to over-fitting. Confirmed outlier samples were excluded from the statistical modelling, however they were included in the testing phase.

As the final step, the data was mean-centered and standardized. Mean-centering is done by subtracting the mean of a variable from every element in the corresponding column. Standardizing consists of dividing each element in a column by the standard deviation. The combination of both methods is beneficial if variables have different ranges of continuous values to allow every variable equal opportunity to influence the final statistical model (Verboven, Hubert and Goos, 2012).

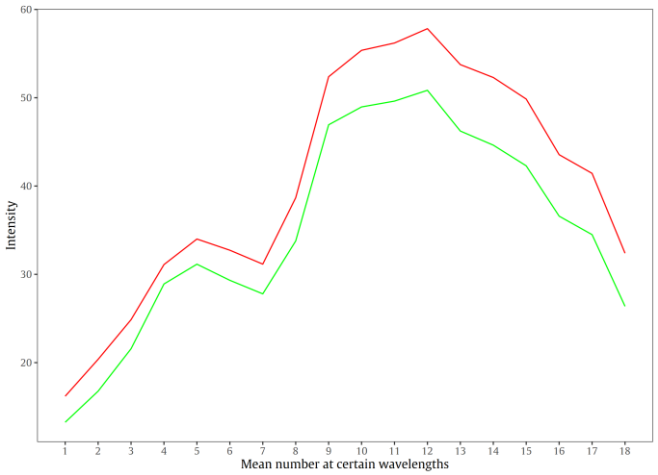



Figure 2 - Typical MSI data collected from fresh and spoiled samples, green: fresh, red: spoiled.

### *2.4 Fourier transform infrared spectroscopy*

Prior to analysis, it was observed that the sample data contained many values above the typical intensity values over 800 $cm^{-1}$. These values correspond to the maximum value which can be obtained from the spectroscopy, however, they do not necessarily contribute to variation. In fact, following further analysis, it was concluded that the noise from this region affected the modelling process negatively. In addition, several previous studies have been found to exclude this region entirely. In the light of this information, variables above 800 $cm^{-1}$ have been excluded from subsequent analyses.

Typical FTIR data collected from fresh and spoiled samples between 4000 and 800 $cm^{-1}$ are shown at *Figure 3*. Major variances in intensity patterns can be seen between 1700-800 $cm^{-1}$ and 3500-3100 $cm^{-1}$ ranges which could be an indicator of spoilage ) (Barth, 2007; Lu and Rasco, 2010; Ellis, Harrigan and Goodacre, 2011; Hernández-Martínez *et al.*, 2014). These regions include:

- Polysaccharide region (~1200-900 $cm^{-1}$);
- Amide III (~1400-1200 $cm^{-1}$);
- Amide II (~1550 $cm^{-1}$);
- Amide I (~1650 $cm^{-1}$);
- Amide A and B (~3300-3070 $cm^{-1}$

Polysaccharide region is of significance because it has been showed that initially, sugar molecules are consumed by the SSO. Eventually, the nitrogenous compounds start to get broken down, creating the vibrations in the amide bands (Nychas *et al.*, 2008). Furthermore, another region of interest was reported to reside above 900 $cm^{-1}$ as the fingerprint region. This region is considered unique for every compound, hence it could be utilised to differentiate

between compounds (Davis and Mauer, 2010; Davis *et al.*, 2010).

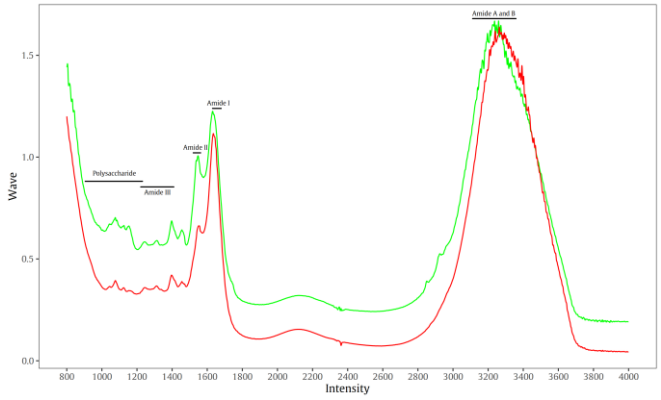



Figure 3 – Typical FTIR data collected from fresh and spoiled samples with regions of interest highlighted. Green: fresh, red: spoiled.

Even though exploratory analyses suggested that 1700-800 and 3500-3100 $cm^{-1}$ bands should be included in the statistical modelling, approaches of previous research were conflicting. Many methodologies of extracting influential wavenumbers were tested in this study as well, the most prominent one being the use of PCA communality values (Ammor, Argyri and Nychas, 2009; Panagou *et al.*, 2011; Vasconcelos, Saraiva and de Almeida, 2014; Saraiva, Vasconcelos and de Almeida, 2017). Communality represents the consistency of a variable by computing the proportion of variance contributed (Abdi and Williams, 2010). Unlike previous studies, this approach in this study could only filter ~10% of the wavenumbers, which was not sufficient. Another common method is to assess variable importance through machine learning algorithms such as random forest, SVMRFE and PLS-R (Kalousis, Prados and Hilario, 2007; Balabin and Smirnov, 2011). However, initial attempts with several machine learning algorithms showed that stable importance values could not be reached. However, it has been shown that it is in fact possible to reach stable values through multiple iterations (Behnamian *et al.*, 2017). With this in mind, a novel approach has been implemented which averaged variable importance values over 50 ANN (single hidden layer, feed-forward) and PLS-R iterations. The results indicated that on average, 1700-850 $cm^{-1}$ and 3400-3200 $cm^{-1}$ bands were the most influential, hence these bands were selected as the variables for statistical modelling.

To further improve the spectral features and enable better analysis, several pre-treatment options were available, frequently utilised ones being:

- Baseline correction;
- Binning;
- Normalisation;
- Smoothing (i.e. Savitzky-Golay, standard normal variate) (Smith, 2011)

While baseline correction could eliminate differences between spectra due to the nature of the infrared penetration of the spectroscopy device, smoothing tends to reduce noise which often weakens the statistical models (Al-Qadiri *et al.*, 2006). Binning reduces the effective spectral resolution by a certain factor via averaging neighbouring variables (Vidal and Amigo, 2012). In other words, binning by a factor of 4 would reduce spectra with 4000 variables to 1000 variables. This not only would improve the computational time required for the statistical modelling, it would also allow certain peaks to be strongly emphasized. Normalisation in the context of spectral data is generally aimed to reduce the differences between each variable by dividing all variables by the most intense band of choice. It has been shown that 1700-1500 $cm^{-1}$ band, which includes both amide I and II, is one of the prime candidates for it (Davis and Mauer, 2010). Baseline corrected, binned (by a factor of 4), normalised (on 1700-1500 $cm^{-1}$) typical FTIR data from fresh and spoiled samples between 1700-850 $cm^{-1}$ and 3400-3200 $cm^{-1}$ can be seen at *Figure 4*. As a side note, several smoothing methods have been experimented with, however, it has been observed that the previous treatments were adequate to obtain necessary clarity in the data. Therefore, no smoothing has been done. As a final step before the statistical modelling, the data was mean-centered and standardized.

Regarding outliers, cook's distance is utilised similar to the procedure with the MSI data. However, cook's distance is not applicable on datasets with more features than the sample count, for this reason, binning with a factor of 10 is applied prior to the test to reduce high dimensionality. 6 strong outliers per sample were identified and excluded from the statistical modelling but included in the testing phase. Weak outliers are ignored to capture more variance.

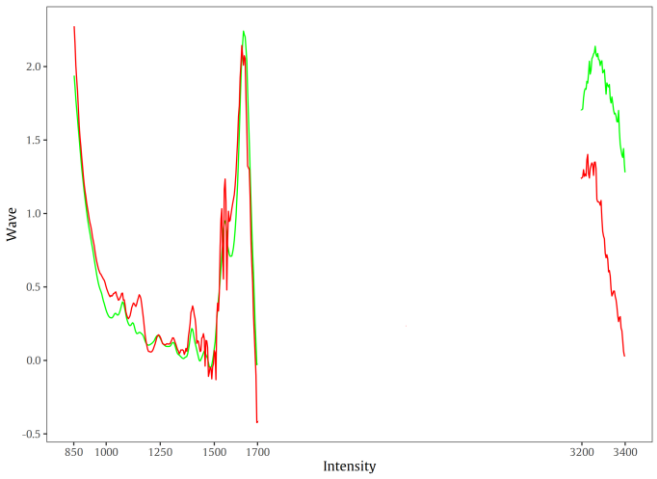



Figure 4 - Pre-processed typical FTIR data collected from fresh and spoiled samples. Green: fresh, red: spoiled.

### *2.5 A closer look at the batches*

It is also important to understand inter-batch variability that could shape the final results. To explore the interaction between the samples in-depth, principal component analysis (PCA) has been done on both MSI and FTIR data. PCA transforms the initial variables (mean intensities in this case) into new small set of variables without losing a great deal of the information from the original data set (Smith, 2002). A common use of PCA in this context is to investigate if there is any clear separation between the samples, which in this case, could give indication of spoilage. TVC values of the sample data have been divided into three categories

(low, med and high) for exploratory purposes and first two principal components (PCs) are plotted for each batch.

PCA results for MSI showed that 99% of the variance could be explained by 5 principal components (PCs) for both batches. Plot for the first two principal components of both batches can be seen at *Figure 5*. Although no clear separation exists for both batches, batch 1 shows some separation while batch 2 shows near none. This could indicate that batch 1 is more suited for statistical modelling since captured variance is often directly correlated with a model's predictive capability. Moreover, PC2 appears to be responsible for the separation more than PC1. In addition, contribution of features (means) to first five PCs (99% variation) were computed. Means 16 and 17 which correspond to 940 and 950 nm were identified as most influential.

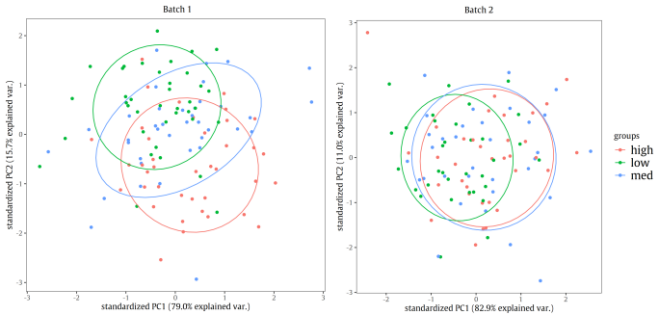



Figure 5 - PCs 1 and 2 plotted for MSI data.

Likewise, results for FTIR showed that 98% of the variance could be explained by 5 PCs for both batches. Plot for the first two PCs can be seen at *Figure 6*. However, unlike MSI data, no separation exists in both batches. Furthermore, possible outliers can be clearly observed, but as previously pointed out, no procedures were carried out to capture more variance in the modelling stage. Similar to MSI data, PC2 appears to be the driving factor for the separation between the samples.

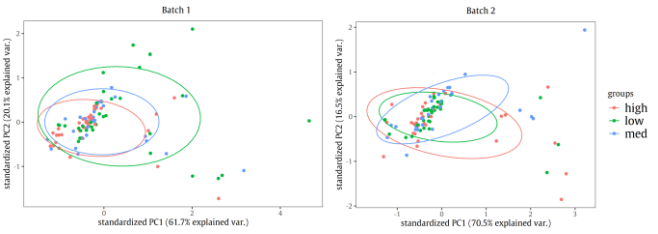



Figure 6 - PCs 1 and 2 plotted for FTIR data.

### *2.6 Statistical modelling*

General outline for the statistical modelling procedure can be seen at *Figure 7*. The modelling and validating have been done in two phases. For the first phase, each batch was tested internally or combined. In other words, models were trained on one part of the batch (or combined) and validated on the remainder. Second phase consisted of training the model on batch 1 and validating on batch 2, or vice versa. This was done mainly to compare the results of both phases to evaluate the effects of inter-batch variability.

"No Free Lunch" is a theorem that have been prominent since the modern era of machine learning. It states that there cannot be a single best model which would work optimally for every problem and dataset (Wolpert, 2002). To put it simply, although some problems tend to favour some machine learning algorithms, at the end of the day, the best result will come from a comparison of different approaches. Therefore, this study features nine different supervised machine learning regression algorithms, namely:

- Partial-least squares (pls) (Geladi and Kowalski, 1986);
- Support vector machine with linear kernel (svmLinear) (Cortes and Vapnik, 1995);
- Support vector machine with radial basis function kernel (svmRadial) (ibid.);
- Random forests (rf) (Breiman, 2001);
- K-nearest neighbours (knn) (Cover and Hart, 1967);
- Principal component regression (pcr) (Jolliffe, 1982);
- Least-angle regression (lars) (Efron *et al.*, 2004);
- Ridge regression (ridge) (Hoerl and Kennard, 1970);
- Artificial neural network (nnet) (Jain, Mao and Mohiuddin, 1996)

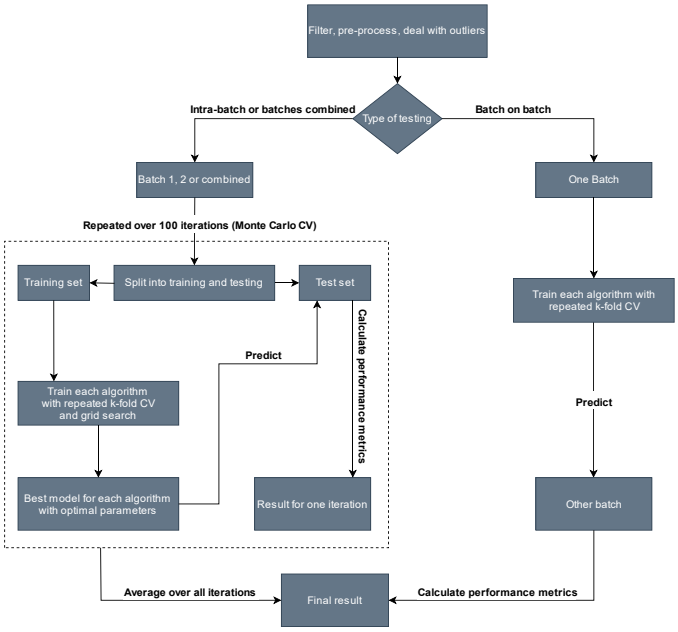



Figure 7 - Outline of the statistical modelling procedure.

For the first phase, the dataset (one batch or both batches combined) is split randomly, while preserving the TVC distribution, into training and test sets with a 70%-30% split. The training set is then trained for each machine learning algorithm, using repeated k-fold cross validation (k=10, repeats=3) and grid search to obtain best performing models with the optimal parameters. Cross validation is useful to avoid over-fitting, which is an instance where the model learns the specifics (i.e. noise) of the training set too well, and consequently, cannot generalize well enough. K-fold cross validation splits the training data into k random groups, trains the model on k-1 groups and evaluates it on the remaining group. This is iterated for each unique group, and for repeated k-fold cross validation, the whole process is repeated for the specified times. Furthermore, cross

validation is reiterated for all possible parameter combinations that were specified in the grid search. Parameter grids that were used can be seen at Appendix B. The final model is selected by comparing the performances of all iterations (Wong, 2015). After the training stage, the final model is validated on the test set to assess overall performance. However, the performance of this method highly depended on the random training/test split that was done at the start. To get an overall appropriate result without this bias, Monte Carlo cross validation is implemented (k=100). This procedure repeats the process outlined above for a number of times with different training and test splits, and averages the performance of all iterations (Xu and Liang, 2001). For the second phase, one batch is trained with k-fold cross validation (k=10, repeats=3) and the best model is validated on the other batch.

### *2.7 Performance metrics*

The performance of the models is evaluated by four different metrics:

- Root mean squared error (RMSE);
- Mean absolute error (MAE);
- R-squared ($R^2$);
- Accuracy

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(Predicted_i - Actual_i)^2}{N}}$$

Equation 1 – Equation for root mean squared error. N: Sample count, $Predicted_i$ : Predicted value at index i resulted after validation, $Actual_i$: Original sample value at index i.

RMSE measures the square root of the average of squared difference between the prediction and the actual values. Even though they are used for different purposes, one can think RMSE and standard deviation as similar measures. For instance, RMSE of 1.5 would mean that 68% of the predictions fall within 1.5 off the actual values, assuming gaussian distribution. RMSE is sensitive to outliers since the difference between the predicted and the actual values are getting squared. It is one of the industry standards for evaluating model performance. Smaller RMSE indicates better performance.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|Predicted_i - Actual_i|$$

Equation 2 – Equation for mean absolute error. N: Sample count, $Predicted_i$: Predicted value at index i, resulted after validation, $Actual_i$: Original sample value at index i.

On the other hand, MAE measures the average of absolute difference between the prediction and the actual values. In other words, if the MAE of a model is 1.5, it can be expected that the average prediction is 1.5 off from the actual value. Similar to RMSE, smaller MAE suggests better performance.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}$$

Equation 3 – Equation for R-squared. N: Sample count, $Y_i$: Original sample value at index i, $\hat{Y}_i$: Fitted value at index i, $\bar{Y}$: Mean value of all values.

$R^2$ is useful to understand how well the independent variables explain the variance in the model. To put it simply, the total difference between the predictions and the actual values are divided by total sum of squares (difference between the prediction/actual slope and average actual values). $R^2$ values are often between 0 and 1 but it can theoretically go to -∞. Negative values indicate a fit worse than assigning the mean average of actual values to all predictions. Values closer to 1 indicate better performance.

$$Accuracy = \frac{Prediction\ count\ within\ 1\ \log10\ cfu}{Prediction\ count}$$

Equation 4 – Equation for accuracy.

Although accuracy is not an applicable metric for regression problems, one can tailor it specifically for the problem. In a similar study, it has been previously considered that TVC predictions within 1 $\log_{10}$ cfu off the actual values are considered accurate (Estelles-Lopez *et al.*, 2017). This approach has been adopted in this study as well.

## 3. Results

### *3.1 Multispectral imaging data results*

The results for the MSI data, consisting of internal testing on batches 1 and 2, and combined, averaged over 100 iterations (Monte Carlo cross validation) are shown at *Figure 8*. For visualization purposes, the models are ranked depending on their average accuracy values. Results showed vast differences between the models, especially a clear contrast can be seen at the bottom three. RMSE values ranged from 1.387 to 0.717, MAE from 1.158 to 0.554, $R^2$ from -12.064 to 0.725 and accuracy from 43.5% to 84.1%. Averaged over all results, svmRadial, rf and knn was ~15%-20% less accurate than the rest. Moreover, while batch 1 attained 73.44% overall accuracy over all models with 0.90 RMSE, batch 2 lagged behind with 65.66% accuracy and 0.99 RMSE. Combined batches showed an in-between result of both batches with 68.55% accuracy and 0.96 RMSE. Highest performance was achieved with artificial neural network with 84.1% accuracy on batch 1, 79.1% on batch 2, and 78.7% on batches combined. Additionally, ridge, lars, pcr, pls and svmLinear performed similarly well with only ~3% behind.

In the same fashion, results for batch-on-batch are shown at *Figure 9*. A clear drop in performance can be observed compared to intra-batch testing, with RMSE values ranging from 1.995 to 1.252, MAE from 1.710 to 0.993, $R^2$ from -23.368 to 0.246 and accuracy from 27% to 56% . Batch 1 appears to be a better predictor on batch 2 with around 24% higher performance on accuracy. In contrast with intra-batch testing, lars, pls and ridge outperformed nnet on average with lars achieving the highest accuracy. Moreover, despite batch 1 outperforming batch 2 as a predictor all

around, $R^2$ values for batch 2 were better, especially for top half of the ranked models. The parameters that were selected from k-fold cross validation for the highest performing method (nnet) on training batch 1 was 1 hidden layer with a decay of 0.1. For training batch 2 (lars) was a fraction of 0.71.

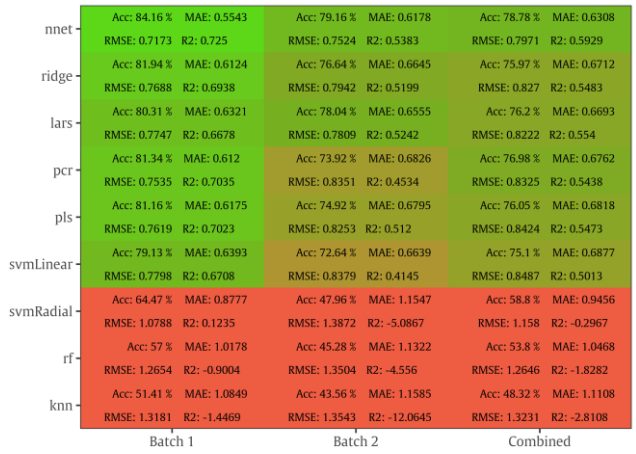



Figure 8 - MSI intra-batch results ranked by validation accuracy.

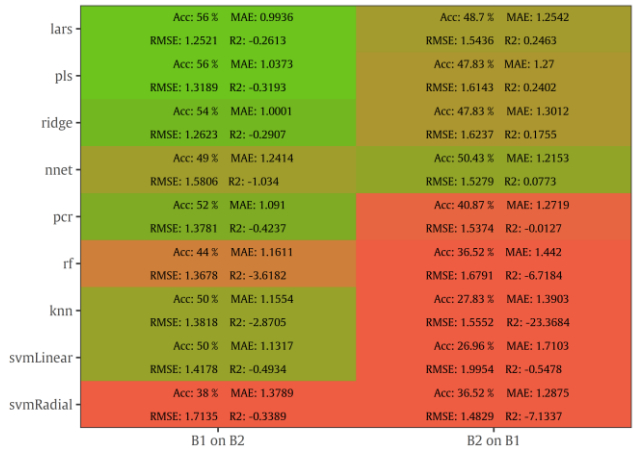



Figure 9 - MSI batch-on-batch results ranked by validation accuracy.

### *3.2 Fourier transform infrared spectrum data results*

The results for the FTIR data, consisting of internal testing on batches 1 and 2, and combined, averaged over 100 Monte Carlo cross validation are shown at *Figure 10*. Performance difference between the models were more compact than the MSI data with RMSE values ranging from 1.536 to 0.857, MAE from 1.164 to 0.669, $R^2$ from -3.129 to 0.5455, and accuracy from 50.0% to 75.9%. Unlike MSI results, batches combined outperformed individual batches by ~5% on accuracy, ~15% on RMSE and MAE, averaged over all models. Furthermore, predictions on batch 2 were significantly more accurate along with a lower error rate than batch 1. Artificial neural network performed best on batch 1 and batches combined with 70.74% and 75.9% accuracies respectively, while lars outperformed nnet by ~13% with 75.2% accuracy and 0.904 RMSE on batch 2 with svmLinear (73.5%, 0.954) a close second. Moreover, nnet, lars and svmLinear outperformed the rest of the models by a decent margin with pls, svmRadial and pcr only catching up on batches combined.

Additionally, batch-on-batch results can be seen at *Figure 11*. In like manner, performances are ranked depending on the average accuracy values. Results differed with previously shown MSI results, with several models outperforming their intra-batch counterparts with RMSE values ranging from 3.924 to 0.8511, MAE from 2.426 to 0.676 and $R^2$ from -8.541 to 0.3538. Training the model on batch 1 and validating on batch 2 outperformed the other way around significantly with ~17% higher accuracy, 55% lower RMSE and 49% lower MAE. In addition, a strong separation can be seen between nnet, lars, svmLinear, pcr and pls compared to the rest of the models for batch 1 on batch 2. Although the highest accuracy was achieved with nnet with 74.26% with training on batch 1 and validating on batch 2, lowest RMSE (0.851) and MAE (0.676) was attained with lars. $R^2$ values showed positive correlation with the overall results, with higher values coming from batch 1 on batch 2 testing. The parameters that were selected from k-fold cross validation for the highest performing algorithm (lars) for training batch 1 was a fraction of 0.17. For training batch 2 (nnet), they were a single hidden layer and a decay of 0.9.


Figure 10 - FTIR intra-batch results ranked by validation accuracy.

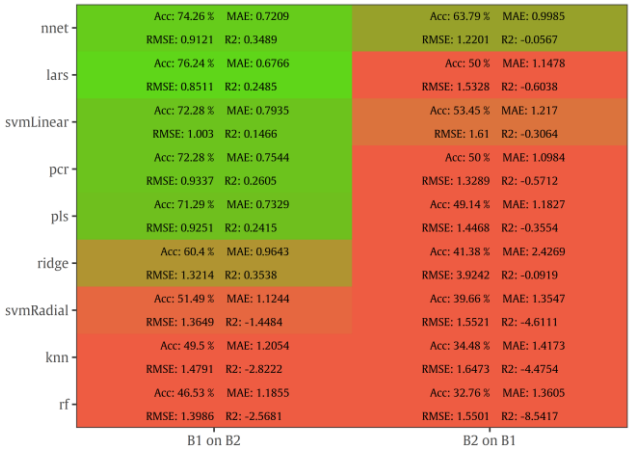



Figure 11 - FTIR batch-on-batch results ranked by validation accuracy.

# 4. Discussion

One of the first things that can be observed from the results above is the effects of inter-batch variability. The performance of the models differs greatly depending on which dataset they are trained at. At first glance, MSI results seems to be superior to FTIR, especially on intra-batch performance for the highest ranked models (80.6% vs 69.3% accuracy). This indicates that MSI could perform very well when the test samples are usually uniform, and the product does not suffer from inter-batch variability over the production line. The reason for the disparity between the platforms could be due to weak outliers which were not been removed from the FTIR batches. Going back to the PCA plots (*Figure 6*), it can be seen that MSI samples form a highly uniform pattern where FTIR has several samples outside of the main cluster. However, as previously pointed out, outliers are often legitimate observations which could give insight into different datasets. In fact, training on batch 1 and testing on batch 2 for FTIR significantly outperformed batch-on-batch performance of MSI (76.2% vs 56%). It is evident that given a quality FTIR dataset with large variability, several models are able to attain respectable prediction performance on an entirely different sample set, which is how a real-world scenario would often look like. Conversely, this was not possible with MSI data in this study. A single additional run with the MSI data for the highest performing models for each batch is plotted by prediction and actual values, which are shown at *Figure 12*. It can be seen that MSI data dramatically mis-predicts most of the samples. This can be due to over-fitting, as the exploratory analysis showed that batch 1 had higher intensity values across the board for similar TVC values. Hence, when trained on batch 1, models tend to over-predict the TVC values on batch 2. In the same manner, training on batch 2 leads to under-predicting batch 1 values. In contrast, plots with the similar approach for the FTIR data are shown at *Figure 13* where it can be observed that predictions and actual TVC values had higher uniformity. This is also true for the model that is trained on batch 2 where the RMSE is closer to the MSI results.

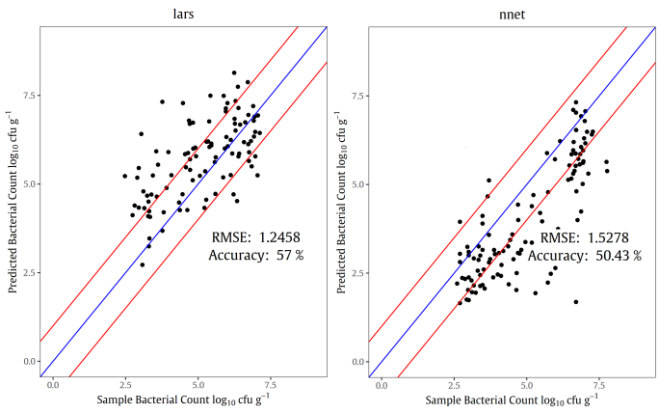



Figure 12 - Plot of predicted vs actual for the highest performing models for MSI batch-on-batch testing. Red lines indicate the boundaries for predictions within 1 $\log_{10}$ cfu.

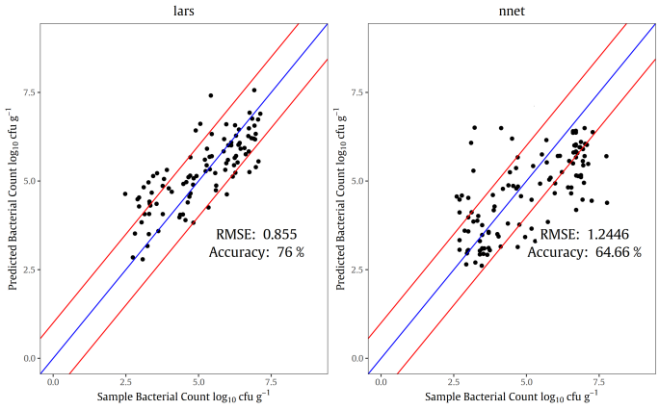



Figure 13 - Plot of predicted vs actual for the highest performing models for FTIR batch-on-batch testing. Red lines indicate the boundaries for predictions within 1 $\log_{10}$ cfu.

Another outcome that can be observed is the correlation between the performance and the sample size. FTIR models tend to get more accurate as the sample size increases (batch 1: 64%, batch 2: 70.5%, combined: 73.4%), which is to be expected since pre-processed data had a total of 274 features and it has been discussed that the ideal training sample size for a set of features is around 10 times the feature set size (Raudys and Jain, 1991). Although sample size does not seem to affect the performance of the models on the MSI data, it could be argued that more samples from different batches would improve the batch-on-batch results as the models had poor generalisation ability.

Regarding individual machine learning algorithm performances, artificial neural network (nnet) performed better overall than other models with least-angle regression (lars) a close second. It should be no surprise that nnet came out on top since most previous similar studies also identified it as the best all-around algorithm. This could be due to its high tolerance to noisy data. On the other hand, the success of lars might be explained by its forward selection behaviour as it incorporates lasso, which modifies regression coefficients in the training phase. This approach is effective in dealing with correlated predictors, which is quite abundant in both datasets. Moreover, limiting the predictor pool tends to reduce over-fitting while increasing the generalising ability of the models (Hesterberg *et al.*, 2008). In addition, lars does not suffer from longer training times as other stepwise selection methods, however, it could be considered as greedy.

Despite their popularity, svmRadial, rf and knn was observed as largely inferior compared to other algorithms in this study. On a closer inspection, it was detected that all three algorithms tend to severely over-fit on both platforms, leading to poor prediction on unseen data. A plot of actual vs predicted for FTIR batch 1 on training (itself) and test data (batch 2) for rf can be seen at *Figure 14*. This could be explained by the low sample size, as these algorithms often thrive on large datasets. Another point of interest was that, contrary to svmRadial's poor performance, its linear counterpart, svmLinear, had decent results for the FTIR data.

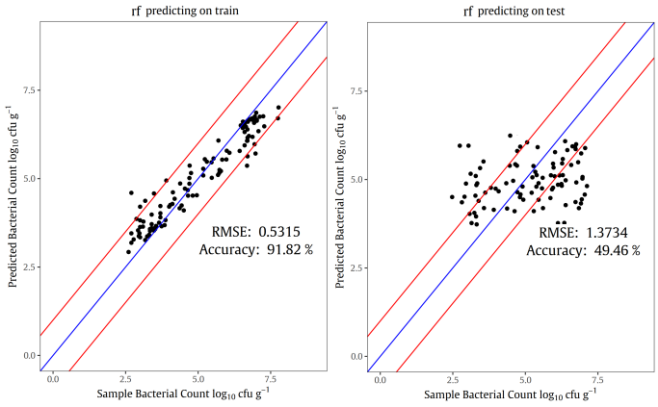



Figure 14 - Demonstration of possible over-fitting for FTIR with Random Forest. First plot is training on batch 1 and testing on the training data. Plot 2 is testing on batch 2.

Even though the results are encouraging, there is definitely room for improvement. It can be argued that the results can only be as good as the experimental data provided. For instance, a bigger sample size would mean models with higher accuracy and better generalisation ability. This is particularly relevant with food products since any environmental factor, such as bacterial contamination, that could affect the biological structure of the material would certainly influence the outcome. Moreover, as previously stated, inter-batch variability was observed to play a big role in the final results. Perhaps training data from multiple unrelated batches could be the solution. Furthermore, even though artificial neural network was identified as the model with highest overall performance, it should be considered that artificial neural networks cover a vast range of different algorithms. In this study, only basic feed-forward, 1-2 hidden layer nnet is utilised. Performance could certainly be improved through different network structures, such as deep convolutional networks. Another improvement could come from feature selection techniques for the FTIR platform. To the best of our knowledge, there is no established approach to extract the most impactful information from the spectra. Although a novel method was used in this study, there is perhaps room for improvement through the use of chemometrics and machine learning.

# 5. Conclusions

The findings of this study showed that both MSI and FTIR data could be an effective instrument for monitoring the spoilage of chicken breast fillets. While data from the MSI platform excelled at predicting similar samples (i.e. same batch), FTIR came through on highly different samples (i.e. different batch). Artificial neural network and least-angle regression were identified as the models with the highest accuracies across the board. Nevertheless, it should be emphasized that the models developed can only be as good as the experimental data supplied, and errors from sensory or microbiological analysis could lead to poorly functioning models. Moreover, a bigger sample set which incorporates different batches would most likely lead to better performance while allowing the ability to employ advanced neural network architectures. Therefore, further research is required for these methods to be eligible on real-world scenarios where the goal is to monitor spoilage on samples from a variety of different storage conditions and production lines.

# Acknowledgements

The Laboratory of Microbiology and Biotechnology of Foods, of the Agricultural University of Athens, Greece is acknowledged for supplying the sample data. Moreover, the study was supervised by Dr Fady Mohareb and Dr Maria Anastasiadi who contributed through their vast experience in the field.

# References

Abdi, H. and Williams, L. J. (2010) 'Principal component analysis - Abdi - 2010 - Wiley Interdisciplinary Reviews: Computational Statistics - Wiley Online Library', *Wiley Interdisciplinary Reviews: …*.

Al-Qadiri, H. M. *et al.* (2006) 'Rapid detection and identification of Pseudomonas aeruginosa and Escherichia coli as pure and mixed cultures in bottled drinking water using fourier transform infrared spectroscopy and multivariate analysis', *Journal of Agricultural and Food Chemistry*. doi: 10.1021/jf0609734.

Ammor, M. S., Argyri, A. and Nychas, G. J. E. (2009) 'Rapid monitoring of the spoilage of minced beef stored under conventionally and active packaging conditions using Fourier transform infrared spectroscopy in tandem with chemometrics', *Meat Science*. Elsevier Ltd, 81(3), pp. 507–514. doi: 10.1016/j.meatsci.2008.10.015.

Argyri, A. A. *et al.* (2010) 'Rapid qualitative and quantitative detection of beef fillets spoilage based on Fourier transform infrared spectroscopy data and artificial neural networks', *Sensors and Actuators, B: Chemical*. Elsevier B.V., 145(1), pp. 146–154. doi: 10.1016/j.snb.2009.11.052.

Balabin, R. M. and Smirnov, S. V. (2011) 'Variable selection in near-infrared spectroscopy: Benchmarking of feature selection methods on biodiesel data', *Analytica Chimica Acta*. doi: 10.1016/j.aca.2011.03.006.

Barth, A. (2007) 'Infrared spectroscopy of proteins', *Biochimica et Biophysica Acta - Bioenergetics*, 1767(9), pp. 1073–1101. doi: 10.1016/j.bbabio.2007.06.004.

Behnamian, A. *et al.* (2017) 'A Systematic Approach for Variable Selection with Random Forests: Achieving Stable Variable Importance Values', *IEEE Geoscience and Remote Sensing Letters*. doi: 10.1109/LGRS.2017.2745049.

Breiman, L. (2001) 'Random Forrest', *Machine Learning*. doi: 10.1023/A:1010933404324.

Carstensen, J. M. and Hansen, J. F. (2003) 'An apparatus and a method of recording an image of an object, Patent family EP1051660', p. 2019.

Cortes, C. and Vapnik, V. (1995) 'Support-Vector Networks', *Machine Learning*. doi: 10.1023/A:1022627411411.

Cover, T. M. and Hart, P. E. (1967) 'Nearest Neighbor Pattern Classification', *IEEE Transactions on Information Theory*. doi: 10.1109/TIT.1967.1053964.

Daalgard, P. (1995) 'Qualitative and quantitative characterization of spoilage bacteria from packed fish', *International journal of food microbiology*, 26(94), pp. 319–333. Available at: http://coli.usal.es/web/alteracion/unidades/labV/SSSP/53Docu/QualitativeQuantitativeCharacterizationSpoilage.pdf.

Davis, R. *et al.* (2010) 'Detection of E. coli O157:H7 from ground beef using fourier transform infrared (FT-IR) spectroscopy and chemometrics', *Journal of Food Science*. doi: 10.1111/j.1750-3841.2010.01686.x.

Davis, R. and Mauer, L. (2010) 'Fourier transform infrared (FT-IR) spectroscopy: a rapid tool for detection and analysis of foodborne pathogenic bacteria', *Current Research, Technology and Education Topics in Applied Microbiology and Microbial Biotechnology. A. Méndez-Vilas (Ed.)*, (1), pp. 1582–1594. Available at: http://www.formatex.info/microbiology2/1582-1594.pdf.

Dissing, B. S. *et al.* (2013) 'Using Multispectral Imaging for Spoilage

Detection of Pork Meat', *Food and Bioprocess Technology*, 6(9), pp. 2268–2279. doi: 10.1007/s11947-012-0886-6.

Efron, B. *et al.* (2004) 'Least angle regression', *Annals of Statistics*. doi: 10.1214/009053604000000067.

Ellis, D. I. and Broadhurst, D. (2002) 'Rapid and quantitative detection of the microbial spoilage of meat by Fourier transform infrared spectroscopy and machine learning', *Applied and Environmental Microbiology*, 68(6), pp. 2822–2828. doi: 10.1128/AEM.68.6.2822.

Ellis, D. I., Broadhurst, D. and Goodacre, R. (2004) 'Rapid and quantitative detection of the microbial spoilage of beef by Fourier transform infrared spectroscopy and machine learning', *Analytica Chimica Acta*. doi: 10.1016/j.aca.2004.03.060.

Ellis, D. I., Harrigan, G. G. and Goodacre, R. (2011) 'Metabolic Fingerprinting with Fourier Transform Infrared Spectroscopy', in *Metabolic Profiling: Its Role in Biomarker Discovery and Gene Function Analysis*. doi: 10.1007/978-1-4615-0333-0_7.

Estelles-Lopez, L. *et al.* (2017) 'An automated ranking platform for machine learning regression models for meat spoilage prediction using multi-spectral imaging and metabolic profiling', *Food Research International*. Elsevier, 99(May), pp. 206–215. doi: 10.1016/j.foodres.2017.05.013.

Geladi, P. and Kowalski, B. R. (1986) 'Partial least-squares regression: a tutorial', *Analytica Chimica Acta*. doi: 10.1016/0003-2670(86)80028-9.

Girolami, A. *et al.* (2013) 'Measurement of meat color using a computer vision system', *Meat Science*. Elsevier Ltd, 93(1), pp. 111–118. doi: 10.1016/j.meatsci.2012.08.010.

Hernández-Martínez, M. *et al.* (2014) 'Application of MIR-FTIR spectroscopy and chemometrics to the rapid prediction of fish fillet quality', *CYTA - Journal of Food*. doi: 10.1080/19476337.2014.889213.

Hesterberg, T. *et al.* (2008) 'Least angle and ℓ 1 penalized regression: A review', *Statistics Surveys*. doi: 10.1214/08-SS035.

Hoerl, A. E. and Kennard, R. W. (1970) 'Ridge Regression: Biased Estimation for Nonorthogonal Problems', *Technometrics*. doi: 10.1080/00401706.1970.10488634.

Huang, Q. *et al.* (2015) 'Non-destructively sensing pork's freshness indicator using near infrared multispectral imaging technique', *Journal of Food Engineering*. Elsevier Ltd, 154, pp. 69–75. doi: 10.1016/j.jfoodeng.2015.01.006.

Jain, A. K., Mao, J. and Mohiuddin, K. M. (1996) 'Artificial neural networks: A tutorial', *Computer*. doi: 10.1109/2.485891.

Jolliffe, I. T. (1982) 'A Note on the Use of Principal Components in Regression', *Applied Statistics*. doi: 10.2307/2348005.

Kalousis, A., Prados, J. and Hilario, M. (2007) 'Stability of feature selection algorithms: A study on high-dimensional spaces', *Knowledge and Information Systems*. doi: 10.1007/s10115-006-0040-8.

Lasch, P. and Naumann, D. (2015) 'Infrared Spectroscopy in Microbiology', *Encyclopedia of Analytical Chemistry*, pp. 1–32. doi: 10.1002/9780470027318.a0117.pub2.

Lu, X. and Rasco, B. (2010) 'Investigating Food Spoilage and Pathogenic Microorganisms by Mid-Infrared Spectroscopy', in *Handbook of Vibrational Spectroscopy*. doi: 10.1002/0470027320.s8967.

MAGDELAINE, P., SPIESS, M. P. and VALCESCHINI, E. (2008) 'Poultry meat consumption trends in Europe', *World's Poultry Science Journal*. doi: 10.1017/s0043933907001717.

Mohareb, F. *et al.* (2016) 'Ensemble-based support vector machine classifiers as an efficient tool for quality assessment of beef fillets from electronic nose data', *Analytical Methods*. Royal Society of Chemistry, 8(18), pp. 3711–3721. doi: 10.1039/c6ay00147e.

Nychas, G.-J. E. *et al.* (2008) 'Meat spoilage during distribution.', *Meat science*, 78(1–2), pp. 77–89. doi: 10.1016/j.meatsci.2007.06.020.

Nychas, G.-J. E., Marshal, D. L. and Sofos, J. N. (2007) 'Meat, Poultry, and Seafood', *Food Microbiology: Fundamentals and Frontiers*.

Panagou, E. Z. *et al.* (2011) 'A comparison of artificial neural networks and partial least squares modelling for the rapid detection of the microbial spoilage of beef fillets based on Fourier transform infrared spectral fingerprints', *Food Microbiology*. Elsevier Ltd, 28(4), pp. 782–790. doi: 10.1016/j.fm.2010.05.014.

Papadopoulou, O. *et al.* (2011) 'Contribution of Fourier transform infrared (FTIR) spectroscopy data on the quantitative determination of minced pork meat spoilage', *Food Research International*. Elsevier Ltd, 44(10), pp. 3264–3271. doi: 10.1016/j.foodres.2011.09.012.

Pedersen, D. K. *et al.* (2003) 'Early prediction of water-holding capacity in meat by multivariate vibrational spectroscopy', *Meat Science*. doi: 10.1016/S0309-1740(02)00251-6.

R Foundation for Statistical Computing. (2018) *R: a Language and Environment for Statistical Computing*., *http://www.R-project.org/*.

Raudys, S. J. and Jain, A. K. (1991) 'Small Sample Size Effects in Statistical Pattern Recognition: Recommendations for Practitioners', *IEEE Transactions on Pattern Analysis and Machine Intelligence*. doi: 10.1109/34.75512.

Rodriguez-Saona, L. E. *et al.* (2001) 'Rapid detection and identification of bacterial strains by Fourier transform near-infrared spectroscopy', *Journal of Agricultural and Food Chemistry*, 49(2), pp. 574–579. doi: 10.1021/jf000776j.

Saraiva, C., Vasconcelos, H. and de Almeida, J. M. M. M. (2017) 'A chemometrics approach applied to Fourier transform infrared spectroscopy (FTIR) for monitoring the spoilage of fresh salmon (Salmo salar) stored under modified atmospheres', *International Journal of Food Microbiology*. Elsevier B.V., 241, pp. 331–339. doi: 10.1016/j.ijfoodmicro.2016.10.038.

Smith, B. C. (2011) *Fundamentals of Fourier Transform Infrared Spectroscopy*, *Fundamentals of Fourier Transform Infrared Spectroscopy*. doi: 10.1201/b10777.

Smith, L. I. (2002) 'A tutorial on Principal Components Analysis Introduction', *Statistics*.

Stevens, J. P. (1984) 'Outliers and influential data points in regression analysis', *Psychological Bulletin*. doi: 10.1037/0033-2909.95.2.334.

Vasconcelos, H., Saraiva, C. and de Almeida, J. M. M. M. (2014) 'Evaluation of the Spoilage of Raw Chicken Breast Fillets Using Fourier Transform Infrared Spectroscopy in Tandem with Chemometrics', *Food and Bioprocess Technology*. doi: 10.1007/s11947-014-1277-y.

Verboven, S., Hubert, M. and Goos, P. (2012) 'Robust preprocessing and model selection for spectral data', *Journal of Chemometrics*. doi: 10.1002/cem.2446.

Vidal, M. and Amigo, J. M. (2012) 'Pre-processing of hyperspectral images. Essential steps before image analysis', *Chemometrics and Intelligent Laboratory Systems*. doi: 10.1016/j.chemolab.2012.05.009.

Van Wezemael, L. *et al.* (2010) 'Consumer perceptions of beef healthiness: Results from a qualitative study in four European countries', *BMC Public Health*, 10. doi: 10.1186/1471-2458-10-342.

Williams, C. K. I. (2003) 'Learning With Kernels: Support Vector Machines, Regularization, Optimization, and Beyond', *Journal of the American Statistical Association*. doi: 10.1198/jasa.2003.s269.

Wolpert, D. H. (2002) 'The Supervised Learning No-Free-Lunch Theorems', in *Soft Computing and Industry*. doi: 10.1007/978-1-4471-0123-9_3.

Wong, T. T. (2015) 'Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation', *Pattern Recognition*. Elsevier, 48(9), pp. 2839–2846. doi: 10.1016/j.patcog.2015.03.009.

Xu, Q. S. and Liang, Y. Z. (2001) 'Monte Carlo cross validation', *Chemometrics and Intelligent Laboratory Systems*, 56(1), pp. 1–11. doi: 10.1016/S0169-7439(00)00122-2.

# Appendix A

## *R packages*

**Cairo** (Simon Urbanek and Jeffrey Horner (2019). Cairo: R Graphics Device using Cairo Graphics Library for Creating High-Quality Bitmap (PNG, JPEG, TIFF), Vector (PDF, SVG, PostScript) and Display (X11 and Win32) Output. R package versio n 1.5-10. https://CRAN.R-project.org/package=Cairo)

**caret** (Max Kuhn. Contributions from Jed Wing, Steve Weston, Andr e Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary M ayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald L escarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, Can Candan and Tyler Hunt. (2019). caret: Classification and Regression Training. R package version 6.0-84. https://CRAN.R-project.org/package=caret)

**caretEnsemble** (Zachary A. Deane-Mayer and Jared E. Knowles (201 6). caretEnsemble: Ensembles of Caret Models. R package version 2 .0.0. https://CRAN.R-project.org/package=caretEnsemble)

**corrplot** (Taiyun Wei and Viliam Simko (2017). R package "corrplot ": Visualization of a Correlation Matrix (Version 0.84). Available fro m https://github.com/taiyun/corrplot)

**doParallel** (Microsoft Corporation and Steve Weston (2018). doPara llel: Foreach Parallel Adaptor for the 'parallel' Package. R package v ersion 1.0.14. https://CRAN.R-project.org/package=doParallel)

**dplyr** (Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2019). dplyr: A Grammar of Data Manipulation. R package version 0.8.3. https://CRAN.R-project.org/package=dplyr)

**FactoMiner** (Sebastien Le, Julie Josse, Francois Husson (2008). Facto MineR: An R Package for Multivariate Analysis. Journal of Statistica l Software, 25(1), 1-18. 10.18637/jss. v025.i01)

**ggplot2** (H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.)

**hyperSpecJSS** (Claudia Beleites and Valter Sergo: `hyperSpec: a pac kage to handle hyperspectral data sets in R', R package version 0.99 -20180627. http://hyperspec.r-forge.r-project.org)

**MLmetrics** (Yachen Yan (2016). MLmetrics: Machine Learning Eval uation Metrics. R package version 1.1.1. https://CRAN.R-project.org /package=MLmetrics)

**stringr** (Hadley Wickham (2019). stringr: Simple, Consistent Wrap pers for Common String Operations. R package version 1.4.0. https: //CRAN.R-project.org/package=stringr)

**prospectr** (Antoine Stevens and Leornardo Ramirez-Lopez (2013). A n introduction to the prospectr package. R package Vignette R pack age version 0.1.3.)

**rms** (Frank E Harrell Jr (2019). rms: Regression Modeling Strategies . R package version 5.1-3.1. https://CRAN.R-project.org/package=rm s)

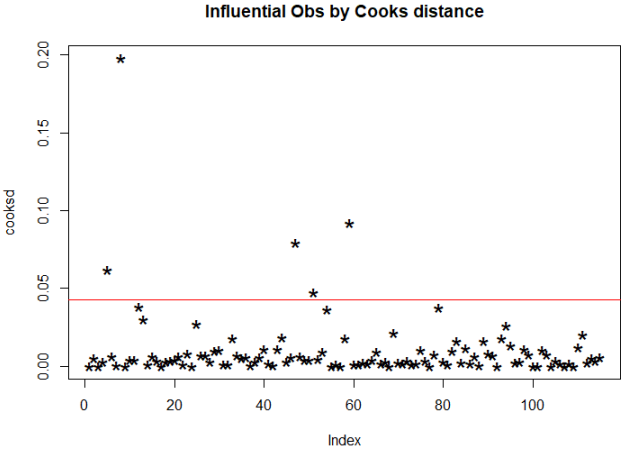
**soil.spec** (Andrew Sila, Tomislav Hengl and Thomas Terhoeven-Urs elmans (2014). soil.spec: Soil Spectroscopy Tools and Reference Mo dels. R package version 2.1.4. https://CRAN.R-project.org/package= soil.spec)

tidyverse (Hadley Wickham (2017). tidyverse: Easily Install and Lo ad the 'Tidyverse'. R package version 1.2.1. https://CRAN.R-project. org/package=tidyverse)

# Appendix B

| | TVC (log cfu $g^{-1}$) | Time (hours) | Temperature (°C) |
|---|---|---|---|
| **B1** | Min.: 2.602<br>1st Qt.: 3.498<br>Median: 5.176<br>3rd Qt.: 6.692<br>Max.: 7.772<br>Std.: 1.568 | Min.: 0<br>1st Qt.: 48<br>Median: 110<br>3rd Qt.: 182<br>Max.: 518<br>Std.: 102<br>Corr. w/ TVC: 0.587 | 0° samples: 35<br>5° samples: 34<br>10° samples: 26<br>15° samples: 20<br>Corr. w/ TVC: 0.438 |
| **B2** | Min.: 2.477<br>1st Qt.: 3.887<br>Median: 5.316<br>3rd Qt.: 6.290<br>Max.: 7.124<br>Std.: 1.348 | Min.: 0<br>1st Qt.: 48<br>Median: 120<br>3rd Qt.: 216<br>Max.: 456<br>Std.: 115<br>Corr. w/ TVC: 0.427 | 0° samples: 34<br>5° samples: 24<br>10° samples: 22<br>15° samples: 20<br>Corr. w/ TVC: 0.387 |

*Descriptive statistics of the batches*




*Demonstration of Cook's distance test on MSI data, batch 1. Red line indicates mean Cook's distance * 4, samples above this line were considered potential outliers*

| Models | Parameters | |
|---|---|---|
| **pls** | ncomp = 1:14 (MSI), 1:40 (FTIR) | |
| **svmLinear** | cost = $2^{0:5}$ | |
| **svmRadial** | cost = $2^{0:5}$ | sigma = $2^{-25, -20, -15, -10, -5, 0}$ |
| **rf** | mtry = 2,4,6,12,18 (MSI), 2,4,6,12,24,48,96,192 (FTIR) | ntree = 500 |
| **knn** | | |
| **pcr** | ncomp = 1:14 (MSI), 1:40 (FTIR) | |
| **ridge** | lambda = $10^{-5:-1}$,1 | |
| **lars** | fraction = 0.1:1 (by 0.01) | |
| **nnet** | Size = 1, 2, maxIteration = 1000 | decay=0.1:1 (by 0.1) |

*Parameters used in CV, every combination is evaluated. 1:15 denotes incremental by 1 (i.e. 1,2,3…,15)*