Classification of Pure and Adulterated Food Samples Using Near-Infrared Spectroscopic Data.

Deirdre Toher*†, Gerard Downey* and Thomas Brendan Murphy†

* Teagasc, The National Food Centre, Ashtown
† Department of Statistics, Trinity College Dublin.

Problem – Pure or Adulterated?

The main aim of food authenticity studies is to detect when foods are not what they claim to be. Honey is defined by the E.U. as a *natural product produced by the European honey bee*, thus any chemical alteration means that the product can no longer be claimed by producers to be honey.

As it is a relatively expensive product to produce and naturally extremely variable, honey is prone to adulteration for economic gain. Indeed cases of honey adulteration have been recorded since Roman times when concentrated grape juice was sometimes added.

False claims may also be made in relation to the origin of the honey, but this study concentrates on attempting to classify samples into pure and adulterated.

The adulteration was completed in the laboratory, using a number of adulterants – fructose:glucose mixtures, beet invert syrup and high fructose corn syrup – and various ratios and weight percentages of these adulterants. The spectra of these samples were then recorded over the wavelength range $1100 - 2498$ nm as seen in Figure 1 below. The similarity of the pure and adulterated spectra is clearly evident, with the pure spectra almost completely masked by the adulterated spectra.

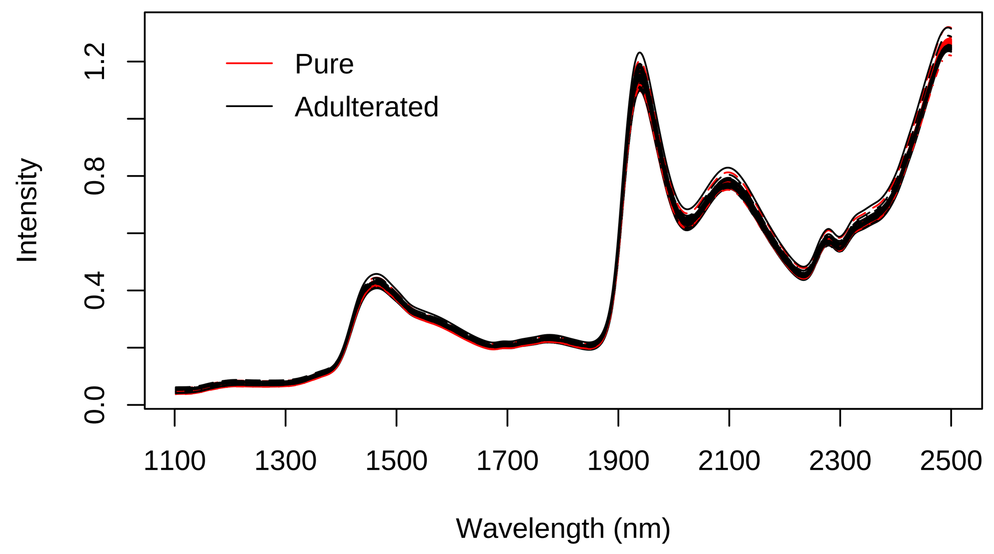

FIGURE 1 *Spectra of Pure and Adulterated Honey*

Dimensionality Reduction

The data span 1100 to 2500 nm, with measurements taken every 2 nm. Adjacent absorption values are highly correlated, therefore dimension reduction is the first issue to be confronted. The technique chosen was that of wavelet analysis.

Wavelet Analysis

Wavelet analysis is used to decompose a spectrum into a series of wavelet coefficients. The coefficients can be used to reconstruct the original spectrum, so no information loss occurs. However, on examining the coefficients produced by the wavelet analysis, it is evident that many are zero or close to zero. Removing such coefficents enables recomposition of the spectra using only fourteen coefficients. The recomposed spectra are then approximations of the original spectra [Figure 2].

Thresholding is used to select the coefficients that contain important information on the structure of the spectrum. Many thresholding techniques have been proposed and the choice of methods is a subjective one.

The Daubechies' wavelet is a consistently reliable type to use and is the default within `wavethresh` [1]. Efficient wavelet analysis methods require that the dimension of the data must be 2^m, where m is an integer. This forces 188 observations to be discarded. The central $2^9 = 512$ observations were chosen – the range $(1290 - 2312)$ nm.

Methods of extending the tails of the spectra in order to have $2^{10} = 1024$ observations are unreliable, with resulting analysis proving problematic – often the associated variance structures are singular.

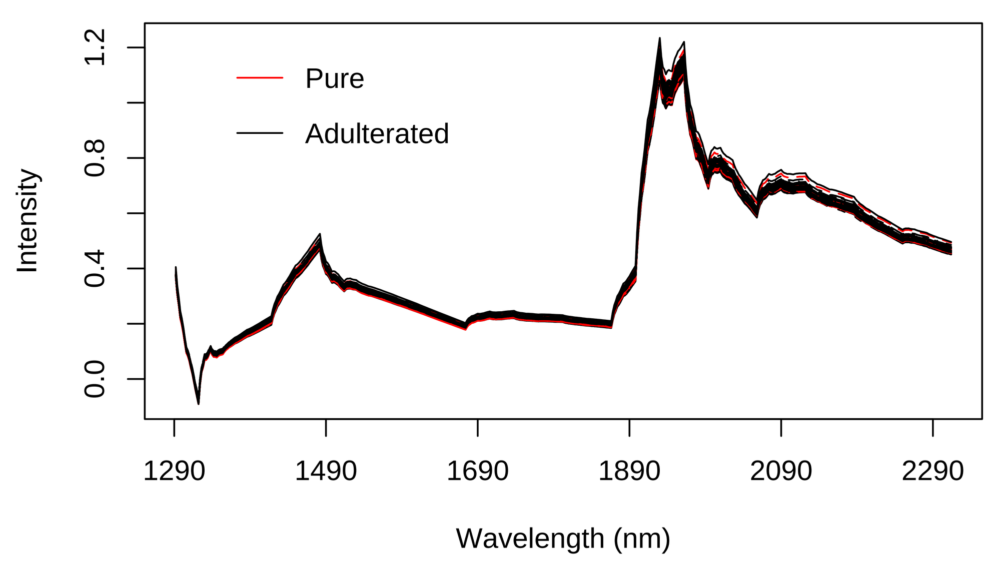

FIGURE 2 *Reconstructed Thresholded Spectra*

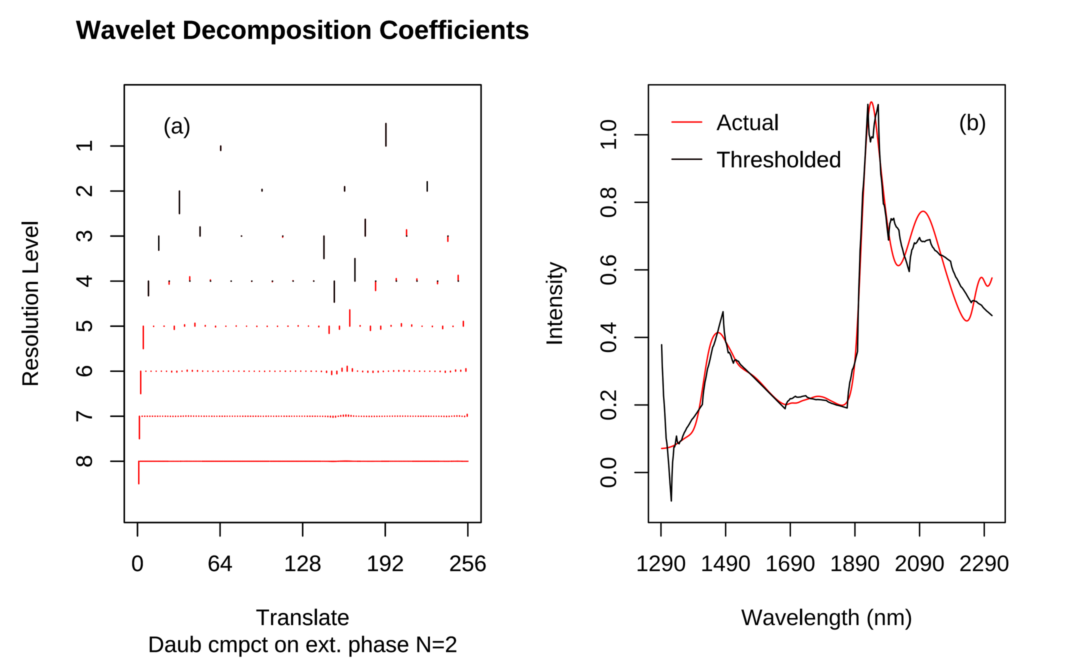

FIGURE 3a *Wavelets Decomposition – thresholded and actual/non-thresholded*

FIGURE 3b *Actual and Reconstructed Thresholded Wavelets*

Figure 3a shows the structure of the wavelet analysis for one sample of pure honey, both in its full and thresholded form, while Figure 3b shows the reconstructed spectrum of the same sample after thresholding together with its actual spectrum.

Classification Techniques

The classification techniques used on this data set are based on Gaussian mixture models; each group is modelled using a Gaussian distribution. The covariance of each of the Gaussian models is structured in a parsimonious manner using constraints. This approach offers the ability to model groups that have distinct volume, shape and orientation properties.

Fraley and Raftery's paper [2] describes a methodological approach to cluster analysis, with specific mention of model-based Discriminant Analysis. Their `mclust` [3] package was used to perform the model-based Discriminant Analysis.

This allows for the possibility of the following models:

TABLE 1: Parametrizations of the covariance matrix Σ_k

Model ID	Decomposition	Distribution
EII	$\Sigma_g = \lambda I$	Spherical
VII	$\Sigma_g = \lambda_g I$	Spherical
EEI	$\Sigma_g = \lambda A$	Diagonal
VEI	$\Sigma_g = \lambda_g A$	Diagonal
EVI	$\Sigma_g = \lambda A_g$	Diagonal
VVI	$\Sigma_g = \lambda_g A_g$	Diagonal
EEE	$\Sigma_g = \lambda D A D^T$	Ellipsoidal
EEV	$\Sigma_g = \lambda D_g A D_g^T$	Ellipsoidal
VEV	$\Sigma_g = \lambda_g D_g A D_g^T$	Ellipsoidal
VVV	$\Sigma_g = \lambda_g D_g A_g D_g^T$	Ellipsoidal

Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) correspond to the EEE and VVV models respectively. The letters of the Model ID represent the volume, shape and orientation of the groups.

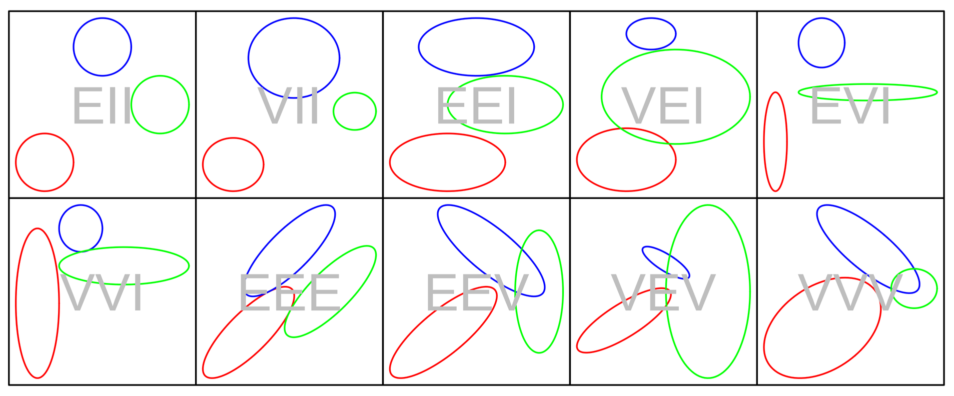

FIGURE 4: *General Shapes of Models*

Model Selection and Verification

400 simulations were performed, each taking a random split of the data to form a training set of 240 samples, with the remaining 238 samples used as a test set. Each training set comprised 79 pure and 161 adulterated honeys (28 using beet invert syrup, 20 using corn syrup and 113 using fructose:glucose mixtures). Test sets comprised 78 pure and 160 adulterated honeys (28 with beet invert syrup, 20 with corn syrup and 112 with fructose:glucose mixtures).

Two types of classification were attempted – to classify into "pure" and "adulterated" and to classify into "pure", "fructose:glucose", "beet invert syrup" and "high fructose corn syrup".

Model selection techniques examined were (a) choosing the model with the best Bayesian Information Criterion value and (b) leave-one-out cross validation.

Classification Results

Selection method: BIC value

Mean Values

		Predicted	
		Adult.	Pure
Actual	Adult.	153.93	**6.07**
	Pure	**8.44**	69.56

Misclassification range
Pure as adulterated: **2–22** Adulterated as pure: **0–14**
Total misclassification: 4–22

Model Selected and Frequency:
EEE/111; EEV/108; VEV/164; VVV/17

Mean Values

		Predicted			
		BI	FG	CS	Pure
Actual	BI	23.34	0.04	0.81	**0.82**
	FG	0.02	105.99	0.00	**6.00**
	CS	4.99	0.60	12.45	**1.96**
	Pure	**0.08**	**6.34**	**0.04**	71.55

Misclassification range
Pure as adulterated: **2–11** Adulterated as pure: **2–21**
Total misclassification (pure/adulterated): 9–27
Incorrect type of adulteration: 2–13

Model Selected and Frequency:
VEV/389; VVV/11

Selection method: Cross Validation

Mean Values

		Predicted	
		Adult.	Pure
Actual	Adult.	154.07	**5.95**
	Pure	**8.13**	69.88

Misclassification range
Pure as adulterated: **2–22** Adulterated as pure: **0–12**
Total misclassification: 7–22

Model Selected and Frequency:
EEE/79; EEV/193; VEV/54; VVV/74

Mean Values

		Predicted			
		BI	FG	CS	Pure
Actual	BI	24.63	0.87	1.88	**0.62**
	FG	0.00	108.08	0.07	**3.85**
	CS	3.15	1.17	14.76	**0.92**
	Pure	**0.75**	**9.30**	**0.70**	67.30

Misclassification range
Pure as adulterated: **2–22** Adulterated as pure: **0–18**
Total misclassification (pure/adulterated): 9–23
Incorrect type of adulteration: 2–15

Model Selected and Frequency:
EEE/122; EEV/127; VEV/144; VVV/7

Conclusions and Further Work

Wavelet analysis provides dramatic dimension reduction while maintaining the differences between pure and adulterated samples. The ability to identify each type of adulteration is promising. The development of a two stage technique of first classifying samples as pure or adulterated then classifying types of adulterations will be explored. Quantifying the level of adulteration also requires further exploration. These techniques are being developed towards a commerical and regulatory standard, and so must be proven to perform not only on Irish honey but on samples from throughout the world.

The potential for using updating algorithms will also be studied.

Acknowledgements

This work was funded by Teagasc under the Walsh Fellowship Scheme.

References

[1] G. Nason, A. Kovac (1997) and M. Maechler (1999), wavethresh: Software to perform wavelet statistics and transforms, R package version **2.2-8**, (2004)

[2] C. Fraley, A.E. Raftery, Model-based clustering, discriminant analysis, and density estimation, *J. Amer. Stat. Assoc.* **97**, 611–631, (2002) available on-line as Technical Report no. 380, October 2000, http://www.stat.washington.edu/www/research/reports

[3] C. Fraley, A.E. Raftery and R. Wehrens (R-port), mclust: Model-based cluster analysis