Hyperspectral Imaging adulterated honey dataset

Tessa Phillips, Bradley Coleman, Shunji Takano, Waleed Abdulla

August 26, 2021

1 Adulteration Sample preparation

In preparing the samples, we want to adulterate each honey to the desired sugar concentrations of $5\%, 10\%, 25\%, 50\%$. Based on this, we made a generous estimate of how much of each honey to heat, aiming at 37 grams. We also aim to heat 20 grams of sugar syrup for each type of honey being adulterated.

The honey samples and a sugar syrup sample are heated to 40 degrees Celcius in an oven, so the honey is homogenous, as per previous work [2]. This temperature is hot enough to melt and mix the honey of all types but not hot enough to damage the active ingredients, particularly in Manuka honey.

First, a sample weighing seven grams of pure honey is poured into a petri dish. Then we will adulterate the solution to contain sugar concentrations of 5%, 10%, 25%, 50%. We work from least adulterated to most adulterated. The amount of sugar desired is calculated based on equation 1. W refers to the total weight of the honey and sugar solution, C_E refers to the existing concentration of sugar in the solution, C_D is the desired concentration of sugar of the final solution, and finally, W_S is the desired weight of sugar to add to the solution to achieve the desired concentration.

$$W_S = ((W * (1 - C_E)) * (\frac{C_D}{1 - C_D}) - (W * C_E) \tag{1}$$

The sugar is then carefully added to the solution. The weight of sugar added is rarely perfect, as we are working with such small amounts. The actual amount of sugar is determined by weighing the amount added to the sample. The actual concentration is determined by the weight of this new sugar and the existing sugar $(W * C_E)$ divided by the new total weight of the sample. This new calculated concentration is used in the next adulteration step as C_E. Any minor errors do not propagate further by using the actual concentration rather than assuming it is close enough to the desired concentration.

Once the adulteration has occurred, seven grams of the solution is poured into a small petri dish for capturing. The actual concentration of sugar is noted down in our metadata providing the label. The overall solution is then weighed again (minus the container weight) to determine the amount of sugar required for the next adulteration step, repeating the process for each sugar concentration required.

Figure 1 shows the overall process of adulterating and capturing our honey samples. The capture, calibration, segmentation, and preprocesing techniques used are all the same as in the pure honey dataset [4].

2 Imagery Calibration

White balancing is a standard calibration procedure for hyperspectral images, as it ensures all the images are based on a consistent amount of light [1][3]. Equation 2 describes the standard white balancing process for the position (x, y), and wavelength λ. Where $\hat{I}(x, y, \lambda)$ is the calibrated hypercube, $I(x, y, \lambda)$ is the original hypercube, and $D(x, y, \lambda)$ and $W(x, y, \lambda)$ are the dark and white references, the dark reference is found by capturing with the camera cap on, and the white reference is found by capturing the calibration material with our standard lighting system.

$$\hat{I}(x, y, \lambda) = \frac{I(x, y, \lambda) - D(x, y, \lambda)}{W(x, y, \lambda) - D(x, y, \lambda)} \tag{2}$$

The dynamic white balancing method extends from the standard white balancing approach and uses a reference value from each row of the hyperspectral image instead of a separate reference for each pixel in the image. Equation 3 shows the dynamic white balancing process.

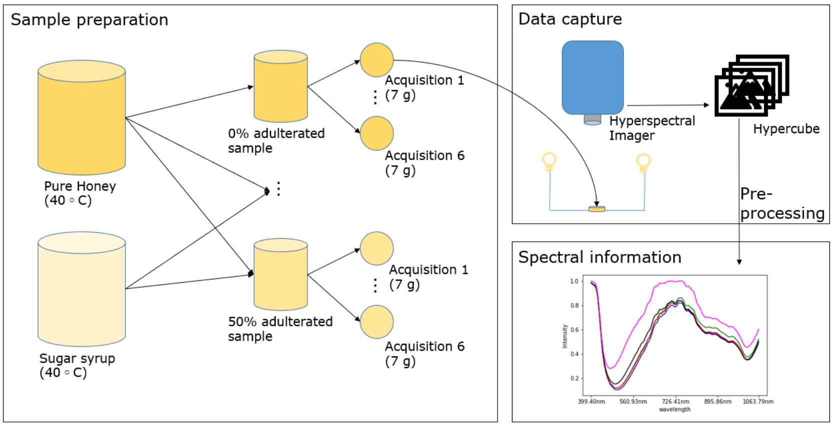

Figure 1: The steps for preparing the hyperspectral imagery samples and extracting spectral information for the adulteration dataset.

$$\hat{I}(x,y,\lambda) = \frac{I(x,y,\lambda) - D(x,y,\lambda)}{W(y,\lambda) - D(x,y,\lambda)} \tag{3}$$

This is important in this work because we are using a push-broom hyperspectral imager so the lighting conditions could change between the rows of the images. The dynamic white balancing approach is what has been used for this dataset [3].

Although in capturing the images the lighting has been kept as consistent as possible, there are some slight inconsistencies due to external lights from the surrounding environment. This needs to be compensated by the proposed white balancing approach.

3 Imagery Segmentation

The segmentation is performed by first cropping the image to remove the background, selecting only the honey sample in the centre of the image as a region of interest. The region of interest is split into 25 segments using a five by five grid [1], as shown in figure 2.

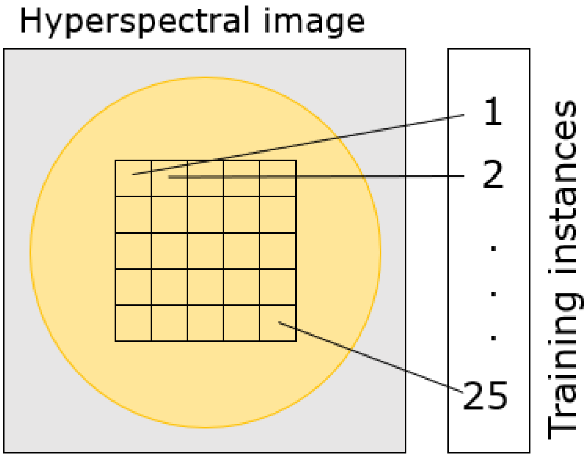

Figure 2: Segmentation of honey imagery through extracting 25 instances from each hyperspectral image of honey. The honey sample yellow circle depicts the honey sample.

The segmentation procedure provides many training examples from each acquisition [3]. The spatial information from the hyperspectral images does not vary significantly as the honey samples are prepared to be homogenous. Yet there are slight perturbations in the data within the grids. In this case, we obtain a lot of training and testing examples sufficient for preparing and testing the predictive models.

The training and testing sets do not split up these segments; rather, each image is either part of the training or testing set. The same approach is followed in the cross-validation.

4 Pre-processing

The pre-processing step follows the segmentation and uses a simple normalisation approach, where the mean of the spectra is forced to be at 0, and the standard deviation to be 1. Other normalisation approaches were also considered; however, this was experimentally found to be the best approach for this dataset [1].

This pre-processing step improves the training convergence, and performance of many machine learning algorithms, such as support vector machines (SVM) and neural networks.

5 Dataset Overview

This dataset consists of segmented and pre-processed hyperspectral images of adulterated honey samples. Overall, the dataset comprises 12 different honey products from seven different brands with 11 different botanical origins labels. Half of the samples are types of Manuka honey which is a premium NZ honey type, and the other half are various types of other NZ honey. Table 1 shows the makeup of the dataset from these different kinds of honey. In creating the dataset, we sampled and captured images of all the honey at each sugar concentration; however, some mixtures were of low quality, and we could not include the images in the final dataset. The full dataset of honey samples contains 8675 total instances. Each instance is the result of spatial segmentation of the hyperspectral imagery. It contains 128 features - representing the spectral wavelengths of the hyperspectral camera. Table 1 represents the number of training and testing examples available. Each sample contains 25 examples following segmentation.

Brand	**Class**	**Adulteration Concentration**					
		0%	**5%**	**10%**	**25%**	**50%**	**Sum**
C1	Clover	150	150	300	300	300	1200
C10	MultiFloral	150	150			150	450
	ManukaUMF5		150	150	150	150	600
	ManukaUMF15		150	150	150	150	600
	ManukaUMF20		150	150	150	150	600
C4	ManukaUMF10		150	150	150	125	575
C5	ManukaBlend		150		150	150	450
C7	BorageField	150	150	150	150	150	750
	Kamahi	150	150	150	150	150	750
	Rewarewa	150	150	150			450
C8	ManukaBlend	150	150	150	150	150	750
C9	Manuka	150	300	300	300	300	1350

Table 1: Overall make-up of the adulterated honey dataset from each brand and botanical origins label of honey.

5.1 Attribute Description

The features from the hyperspectral images are the working wavelengths of the hyperspectral camera. Several additional attributes have been considered and can be useful for splitting the dataset into training and testing sets, as well as testing the generalisation ability of the algorithms.

The ‘Brand’ of honey represents the manufacturer that has supplied the honey. This attribute is included because it can be useful to test if a system can classify all honey types within a brand against each other. It can also be useful when developing general systems to check if we can exclude a brand from the training set and still have a good performance with the testing set. The brands have been anonymised for confidentiality reasons, the brand labels have been renamed as $C1, C2, ..., C11$.

The ‘Acquisition’ attribute represents the different sampling of images for the same type and brand of honey. As portrayed in figure 1, for each unique jar of honey, there have been six samples taken and captured by the hyperspectral imaging system. Each image captured is numbered with an acquisition number between one and six. This attribute allows us to split the training and testing sets such that we obtain a balanced distribution of all the honey types. This also ensures that we do not have an instance in the testing set that comes from a segment included in the training set. For testing, we use acquisition number six, and for training, we use acquisitions one to five.

The class attribute indicates the class of honey, which is the botanical origin, and the UMF value if it is UMF rated Manuka honey. Botanical origins have a huge impact on the value of honey, where some types are precious such as pure Manuka honey, and others are much more common and not considered as valuable, such as multi-floral honey.

Finally there are two sugar concentration attributes. The first 'concentration' represents the actual concentration of sugar in the sample this could be for example 5.001. This attribute is used for regression type algorithms as it has the exact concentration of sugar that we have adulterated the pure honey sample with. The second attribute is 'concentration_class' this represents the class grouping of the adulterated sample as either '0', '5', '10', '25', or '50'. This attribute is used for classification problems to detect what group the adulterated honey should fit into.

References

[1] A. Noviyanto, "Honey botanical origin classification using hyperspectral imaging and machine learning," Ph.D. dissertation, The University of Auckland, 2018.

[2] A. Noviyanto and W. H. Abdulla, "Honey dataset standard using hyperspectral imaging for machine learning problems," in *2017 25th European Signal Processing Conference (EUSIPCO)*. IEEE, 2017, pp. 473–477.

[3] ——, "Segmentation and calibration of hyperspectral imaging for honey analysis," *Computers and Electronics in Agriculture*, vol. 159, pp. 129–139, 2019.

[4] T. Phillips, A. Noviyanto, and W. Abdulla, "Hyperspectral imaging honey database," 4 2020, *Available at https://figshare.com/s/25afe30ff531b8f1e65f*. [Online]. Available: https://figshare.com/s/25afe30ff531b8f1e65f