Detecting Admixture to Mango Purée of the Alphonso Cultivar (*Mangifera indica* L. cv. Alphonso) by ^1H -NMR Spectroscopy

Claas Strecker¹ · Victor Ara¹

Received: 18 March 2021 / Accepted: 9 August 2021 / Published online: 25 August 2021
© The Author(s) 2021, corrected publication 2021

Abstract

Food authenticity is becoming increasingly important but challenges existing analytical methods. In this study, we analyze the mango cultivar Alphonso with regard to authenticity using ^1H -NMR spectroscopy. This cultivar has been termed “the king of mangoes” due to its unique flavor. Regarding its metabolites however, little is known about unique constellations that allow for differentiation of the Alphonso cultivar. We find that the Alphonso cultivar is distinguished by high levels of niacin, trigonelline, and histidine but features relatively low levels of alanine. Furthermore, we develop a model based on the local outlier factor algorithm that effectively detects admixture of non-Alphonso cultivars to Alphonso purée. This task is highly challenging because we identified no metabolites that are unique or uniquely absent in the Alphonso cultivar compared to other mango cultivars analyzed in this study. Our model shows promising results on a test set: Admixtures consisting of 35% non-Alphonso and 65% Alphonso mango purée were uncovered with a sensitivity of 88%. At the same time, our model verified Alphonso samples with a good specificity of 86%.

Keywords Food authenticity · Mango · *Mangifera indica* L. cv. Alphonso · NMR spectroscopy · Anomaly detection

Introduction

Food fraud is no newly emerging phenomenon (Spink and Moyer, 2011; Hong et al., 2017). Yet, food authenticity is becoming increasingly important due to increased awareness of possible health implications resulting from fraud and increased customer appreciation for authentic products (Spink and Moyer, 2011; Hong et al., 2017). However, existing analytical methods for monitoring of food authenticity are challenged by a growing product diversity (Spink and Moyer, 2011; Esslinger et al., 2014).

For authenticity testing, certain analytical techniques have proven particularly useful, e.g., isotopic analysis, DNA-based techniques, and multi-parameter techniques (termed “profiling” in targeted or “fingerprinting” in untargeted mode) (Hong et al., 2017). However, no single approach is able to detect all types of fraud. Of these approaches, DNA-based methods offer the most astonishing sensitivity. For example, Chen et al. were able to detect admixture of

donkey to cattle meat at admixture proportions as low as 0.002% by using a DNA-based method (Chen et al., 2020). Nonetheless, in case of water-based products such as juices, DNA-based methodologies would even fail to detect a very simple yet common form of falsification: water addition. In contrast, measuring the products $\delta^{18}\text{O}$ value by isotope-ratio mass spectrometry might easily reveal such fraud. Multi-parameter approaches offer an efficient way of detecting the manifold of different fraud schemes commonly encountered in reality because balancing many parameters at same time is inherently difficult for falsifiers (McGrath et al., 2018). To save resources, analytical methods that allow retrieving multiple parameters simultaneously are particularly useful for the multi-parameter approach (Esslinger et al., 2014). This explains the popularity of NMR spectroscopy in food profiling and fingerprinting (Esslinger et al., 2014). Additionally, NMR spectroscopy offers a high reproducibility and low requirements for sample preparation (Esslinger et al., 2014).

Hundreds of mango cultivars are known, but the Alphonso cultivar has been termed “the king of mangoes” due to its unique flavor and long shelf life (Deshpande et al., 2017). It is the main cultivar produced in India (Ferrier et al., 2012). The global market for mangoes is growing: In the period from 1998 to 2010, mango consumption per

✉ Claas Strecker
c.strecker@chelab.de

¹ Chelab Dr. V. Ara GmbH & Co. KG, Carl-Zeiss-Straße 16,
30966 Hemmingen, Germany

capita increased by about 50% in the USA (Ferrier et al., 2012). Indian mangoes take a premium niche in this market. For example, the US import price for Indian mangoes was US\$4.2/kg in 2009 compared to an average of US\$0.8/kg for other origins (Ferrier et al., 2012). In the same year, wholesale price of Alphonso mangoes in India ranged from 0.5 to US\$1.6/kg, while the price of the Kesar cultivar ranged only from 0.4 to US\$0.6/kg (Ferrier et al., 2012). As a result of this commercial relevance, the Alphonso cultivar has already been the subject of investigations, i.e., regarding its transcriptome (Deshpande et al., 2017), its volatile components (Pandit et al., 2009), and its fatty acids (Bandyopadhyay and Gholap, 1973). These investigations have mainly focused on changes occurring during ripening. For example, Bandyopadhyay et al. showed that linoleic acid levels decrease but linolenic acid levels increase during ripening of Alphonso mangoes (Bandyopadhyay and Gholap, 1973). Pandit et al. found that ripeness is associated with the appearance of distinct lactones and furanones in the volatile spectrum of Alphonso mangoes (Pandit et al., 2009).

Several studies have investigated mangoes by NMR spectroscopy (Gil et al., 2000; Duarte et al., 2005; Koda et al., 2012; Ryu et al., 2017). However, none of these studies has investigated the Alphonso cultivar or developed a model that is suited to detect admixture of other mango cultivars. In this study, we investigate the Alphonso cultivar by ¹H-NMR spectroscopy and develop a model based on the local outlier factor algorithm that effectively detects admixture of non-Alphonso cultivars to Alphonso purée.

Materials and Methods

Chemicals

Deuterium oxide (99.9% containing 0.05 wt% 3-(trimethylsilyl)propionic-2,2,3,3-d₄ acid (sodium salt)), sodium azide (> 99.5%), nicotinic acid (> 98%), and trigonelline hydrochloride (analytical standard) were purchased from Sigma-Aldrich. Sodium dihydrogen phosphate monohydrate was purchased from Merck.

Sample Preparation

Mango purée (6.8 g) was centrifuged for 10 min at 13,000 rpm. The supernatant was transferred to a separate tube, and 2.0 mL of distilled water was added to the precipitate. The tube containing the precipitate was shaken vigorously and centrifuged for 10 min at 13,000 rpm. The supernatant was combined with the supernatant from the previous step. Combined supernatants were made up to a volume of 8 mL with distilled water. Subsequently, pH was adjusted to 2.90 ± 0.05 using 1.0 M hydrochloric acid. Then,

1.0 mL of buffer solution (1.67 M sodium dihydrogen phosphate + 0.05% sodium azide at pH 2.90) was added, and the solution was made up to a volume of 10.0 mL with distilled water. A proportion of this sample was centrifuged for 10 min at 13,000 rpm. Supernatant (675 µL) and deuterium oxide (75 µL) were mixed. Of this mix, 600 µL was transferred into an NMR tube.

For the training set, ten Alphonso samples were prepared in this manner and subsequently analyzed as described in the next section. Furthermore, 13 samples of non-Alphonso cultivars were prepared for comparison to the training set. These 13 samples consisted of 6 different cultivars: Totapuri (3 ×), Kesar (3 ×), Sindura (3 ×), Chato de Ica (2 ×), Nam Dok Mai (1 ×), and Yai (1 ×). For the test set, seven additional Alphonso samples were prepared. Moreover, 26 self-mixed admixtures were generated. For this, two of the Alphonso samples belonging to the test set were randomly assigned to each of the 13 non-Alphonso samples. The Alphonso sample was then blended with 35% of the non-Alphonso sample.

Data Acquisition

All spectra were acquired on a Bruker Ascend 400 MHz spectrometer (Bruker Biospin, Rheinstetten, Germany) using the noesygppr1d pulse sequence for acquisition of water-suppressed ¹H-NMR spectra. Spectra were recorded at 301.8 K, with 256 scans, 131,072 complex data points, an acquisition time of 7.97 s, a D1 delay of 4 s, and a spectral width of 20.545 ppm. The 90° pulse length (P1) was determined for each individual sample using the *pulsecal* command. The receiver gain was fixed to 32. For data processing, the FIDs were Fourier transformed with a line-broadening factor of 0.3, baseline corrected, and phased with Topspin 3.2 (Bruker Biospin, Rheinstetten, Germany).

Data Analysis

All data analysis was performed using R 4.0.3 (R Core Team, 2020). The following R libraries were used: *stringr* (Wickham, 2019), *ggplot2* (Wickham, 2016), *ggpattern* (Mike, 2021), *signal* (Signal Developers, 2013), *pracma* (Borchers, 2021), *FSelector* (Romanski et al., 2021), *dplyr* (Wickham et al., 2021), *dfoptim* (Varadhan et al., 2020), and *Rlof* (Hu et al., 2020).

Signal Fitting Processed NMR data was imported to R. For signals of ten metabolites (see below), integrals were retrieved by peak fitting with a generalized Lorentzian function (Schoenberger et al., 2016). The following signals were evaluated (all values in ppm): niacin 8.93 (ddd, H-6), histidine 8.65 (d, H-2), adenosine 8.52 (s, H-2), formic acid 8.28 (s), *p*-digallic acid 7.25 (s, H-2/2'), shikimic acid 6.81

(ddd, H-2), choline 3.18 (s, –OH), succinic acid 2.66 (s, H-2/3), alanine 1.49 (d, –CH₃). See supplemental Fig. S1-3 for an exemplary spectrum of an Alphonso sample in that the signals indicated above are labeled. Additionally, a signal at 9.14 ppm (pseudo-s) that originates from overlapping signals of niacin and trigonelline (both H-2) was fitted as well. To obtain the mere peak area of the H-2 proton of trigonelline, the peak area of this signal was subtracted by the previously established peak area of the H-6 proton of niacin. Furthermore, for some of the signals, namely, those of histidine, adenosine, and formic acid, second-order lag-one differences were calculated and fitted to achieve baseline correction. To minimize degradation of signal-to-noise ratio, this step was combined with Savitzky-Golay smoothening. A filter length that improves signal-to-noise but does not significantly affect quantification was determined by an algorithm that iterates over filter lengths and terminates if any peak intensity in the region of interest decreases by more than 1%. After fitting the transformed signals, the fit was backtransformed using the *diffinv* function. Finally, integrals were divided by the 90° pulse length (P1).

Feature Importance To assess feature importance, the information gain was calculated for all ten metabolites. For this, the 10 Alphonso samples forming the training set and the 13 samples of pure non-Alphonso cultivars were used. The latter were not labeled as their actual cultivar but summarized into a single “non-Alphonso” class. Information gain was calculated using the *information.gain* function from the *FSelector* library. Similarly, we assessed the importance of features for discrimination of the 10 Alphonso samples forming the training set and the 26 samples of self-mixed admixtures.

Model Of the ten Alphonso samples forming the training set, one was discarded because it is a significant outlier (cf. results section and Tab. S2). The remaining nine samples were standardized, and their local outlier factor (LOF) was calculated using the *lof* function from the *Rlof* library with $k=4$. LOFs range from 0.969 to 1.101 with a mean of 1.013 and a standard deviation of 0.046.

For new samples (both self-mixed admixtures and new Alphonso samples forming the test set), standardization was performed using mean and standard deviation derived from the training set. To avoid masking, each new sample was appended individually to the samples forming the training set, and the LOF of the new sample was calculated. New samples were evaluated to be authentic Alphonso samples if their LOF was below 1.194. This cut-off value was derived by taking the maximum LOF from the training set and adding two standard deviations as a margin.

Virtual Admixtures Virtual admixtures were generated by linear combination of previously measured samples. For

this, 40 samples were randomly selected from both the Alphonso group (test set) and non-Alphonso group using the *sample_n* function from the *dplyr* library (allowing for multiple drawings of the same sample). Subsequently, their linear combinations were calculated at different admixture proportions (5, 20, 35, 50, 65, 80, and 95%). Then, these virtual admixtures were evaluated in the same way as real samples (see above). To minimize variance, this procedure was repeated ten times.

Results

Spectral Binning Is Not Always Viable

Many studies investigating food authenticity by ¹H-NMR spectroscopy use spectral binning (Esslinger et al., 2014). Binning is a convenient method that quickly captures a relevant share of the information contained in an NMR spectrum. However, binning provides features that offer little interpretability. Furthermore, binning can deteriorate relevant information in case of overlapping signals. We will discuss such a case in the following paragraph.

In Fig. 1a, a segment of the ¹H-NMR spectrum of an Alphonso mango purée is shown. A broad background signal spans from 8.8 to 7.8 ppm. This broad signal likely originates from an oligomeric species that was not identified. Signals of several metabolites (e.g., histidine, adenosine, and formic acid) fall into this region. Binning would mix information of these signals with information of the broad background signal. Therefore, binning would effectively deteriorate information because the broad background signal shows a relevant variance among different samples. Consequently, we decided to employ a targeted approach in that we perform peak fitting for the signals of interest. For this, we use the generalized Lorentzian function (as introduced by Schoenberger et al. (2016)) to fit peaks of ten metabolites: niacin, trigonelline, histidine, adenosine, formic acid, *p*-digallic acid (as assigned by Koda et al. (2012)) shikimic acid, choline, succinic acid, and alanine. Some form of baseline correction is necessary for signals that fall into the region of the broad background signal discussed above. For this, we pick up ideas from Bernstein et al. and fit the second-order lag-one differences (Bernstein et al., 2013). Calculating the second-order lag-one differences rapidly degrades signal-to-noise ratio. Therefore, we combine this step with the application of a Savitzky-Golay filter. Constant and linear terms drop off for the second derivative. As a result, signals that fall into the region of the broad background signal can now easily be fitted (see Fig. 1b). Subsequently, the fit can be backtransformed (see Fig. 1c). Besides allowing for baseline correction, we anticipate that this methodology will enable us to fit highly overlapped signals in the future because signal

Figure 1 **a** Segment of the NMR spectrum of an Alphonso sample. A broad background signal in this region makes spectral binning unfeasible. The adenosine signal that is fitted in **b** and **c** is highlighted by a grey-dashed box. **b** Second-order lag-one differences for the region of the adenosine signal are shown in black. The baseline is flattened by this transformation. As a result, the signal can now easily be fitted (see grey-dashed line). **c** The fit (grey-dashed line) of the adenosine signal (in black) is shown after back-transformation.

widths are significantly narrower for the second derivative. For completeness, we want to point out that experimental possibilities to eliminate macromolecular peaks exist. The Carr-Purcell-Meiboom-Gill (CPMG) pulse sequence is the most popular experiment for this purpose. However, such experiments can attenuate signals and have a negative impact on quantification (Gowda and Raftery, 2014). Additionally, incorporation of such experiments into routine analytics can be a bottleneck.

All signals of trigonelline are however too heavily overlapped by signals of niacin to be unraveled by this strategy (see Fig. S4 and Fig. S5). However, the H-6 proton of niacin is separated from all signals of trigonelline. Therefore, we extract the peak area of the H-2 proton of trigonelline indirectly: We fit the overlapping H-2 protons of both trigonelline and niacin and subsequently subtract the peak area of the H-6 proton of niacin. Integrals obtained by our peak fitting routine were not conveyed to absolute concentrations. Determination of absolute concentrations by NMR requires additional considerations (e.g., regarding incomplete T_1 relaxation) but is of no meaning to our final model anyway because data is standardized in this model. Because all spectra in this study are acquired with identical acquisition parameters, relative peak areas are however proportional to absolute concentrations and comparable between spectra according to the PULCON principle (pulse length–based concentration determination) (Wider and Dreier, 2006). Consequently, integrals were only corrected by dividing by the 90° pulse length.

Alphonso Cultivar Is Distinguished by High Levels of Niacin

We measured purée samples of 10 Alphonso mangoes and 13 samples of other mango cultivars, namely, Chato de Ica, Kesar, Nam Dok Mai, Sindura, Totapuri, and Yai. In a first step, we assessed the ability of each metabolite to

discriminate Alphonso and non-Alphonso cultivars. For this, we calculated the information gain of each metabolite (see Tab. S1 for all values). The concept of information gain is used in decision tree learning. Its values are bound between 0 and 1 (for a binary classification problem), and higher values indicate higher discriminative capabilities. Niacin shows the highest information gain ($IG = 0.685$), followed by histidine ($IG = 0.539$). In Fig. 2a, levels of these two metabolites are shown as a scatterplot. It is apparent that the Alphonso cultivar is distinguished by high levels of niacin and histidine and can be clearly differentiated from all other analyzed cultivars. Additionally, the Alphonso cultivar also features relatively high levels of trigonelline, formic acid (see Fig. 2b), adenosine, and choline (see Fig. S6a). In contrast, alanine has a distinctly low abundance in the Alphonso cultivar (see Fig. S6b). Two metabolites, succinic and *p*-digallic acid, show no discriminative power ($IG = 0$) and were therefore excluded from subsequent analysis.

A Model for Admixture Detection

Many NMR studies have investigated food authenticity, e.g., regarding product variety (Godelmann et al., 2013) origin (Godelmann et al., 2013; Bachmann et al., 2018) or even production conditions (e.g., organic products or production year) (Godelmann et al., 2013; Ackermann et al., 2019). Most of these studies address outright mislabeling. Frequently however, falsifiers have developed elaborated schemes such as mixing authentic and non-authentic products. This poses a significant challenge for analytical methods. In simple cases, the non-authentic fraction may introduce metabolites that are completely absent in authentic products. NMR spectroscopy has already proven that it can achieve relatively low detection limits in such scenarios. For example, Schmitt et al. were able to detect addition of peanut to walnut powder at levels as low as 4% (Schmitt et al., 2020). In cases in that the above condition is not met,

Figure 2 **a** Scatterplot of niacin and histidine levels. The Alphonso cultivar is distinguished by comparatively high levels of these two metabolites. **b** Scatterplot of formic acid and trigonelline levels. The Alphonso cultivar also features comparatively high levels of these two metabolites.

multivariate data analysis becomes necessary. In a good example of this, Bachmann et al. investigated admixtures of hazelnuts of different origins (Bachmann et al., 2019). As their best result, they were able to detect admixture of 50% non-Georgian to 50% Georgian hazelnuts with an accuracy of 75% (Bachmann et al., 2019). Unsurprisingly, accuracy quickly levels off for lower admixture proportions. Admixture of 40% non-Georgian to 60% Georgian hazelnuts could only be detected with an accuracy of less than 20% (Bachmann et al., 2019). This work clearly highlights how challenging it is to detect admixtures if only a small proportion of non-authentic product is admixed to an authentic product.

Most NMR studies use supervised machine-learning classifiers (e.g., linear discriminant analysis or support vector machines) for multivariate data analysis (Esslinger et al., 2014). Even though these methods are undoubtedly powerful, we think that they are not optimal for detection of admixtures. Supervised machine-learning classifiers only perform well if the training data is representative for real-life test data. This condition is frequently not true for admixtures. It is easy to represent the authentic product under analysis in the training data but much more difficult to represent all relevant admixtures of potentially many non-authentic products. In such a scenario, it can even be expected that falsifiers will adapt their strategies once they learn which types of admixtures can easily be detected (McGrath et al., 2018). As a result, the performance of these models will diminish due to new, elaborated fraud schemes. In our opinion, anomaly detection algorithms represent a better fit to the problem at hands because only the authentic product has to be represented in the training data for these. Similar thoughts have been formulated before but are frequently ignored (Oliveri, 2017).

Two prominent anomaly detection algorithms are one-class support vector machines (as introduced by Schölkopf et al. (1999)) and the local outlier factor (LOF, as introduced by Breunig et al. (2000)). One-class support vector machines feature some favorable attributes (e.g., computational speed). However, they do not feature class probabilities, which significantly lower their interpretability. In contrast, the LOF algorithm returns a factor that allows evaluating how “outlierish” an outlier is. The LOF algorithm is a density-based unsupervised anomaly detection method. Due to its unsupervised nature, the LOF algorithm just returns a factor but does not give any decision rule how to interpret this factor. To construct such a decision rule, we used the ten Alphonso samples that were measured in the previous section as a training set. One of the samples is an outlier and was therefore removed from the training set. This sample features an LOF of 1.590 while the LOFs for the other nine samples range from 0.927 to 1.169. This outlier sample is distinguished by a significantly elevated level of histidine (see Fig. 2a). After removal of this sample, LOFs of the remaining samples range from 0.969 to 1.101 with a mean of 1.013 and a standard deviation of 0.046. We decided to take the maximum LOF as a cut-off and add two standard deviations as a margin. As a result, samples are classified as inliers if their LOF is below 1.194.

This procedure has transformed our model into a semi-supervised anomaly detection model meaning it was only trained on inliers. To test our model, we measured seven new Alphonso samples. Only one of the samples was not correctly classified as an inlier, giving our method a specificity of 86%. Subsequently, we tested the ability of our model to detect admixtures. For this, we self-mixed 26 admixtures

Fig. 3 Confusion matrix for the test set. Non-Alphonso samples are admixtures made up of 35% non-Alphonso cultivars and 65% of Alphonso samples. These admixtures were detected with a sensitivity of 88%

composed of 35% non-Alphonso and 65% Alphonso puree. Of these, our model identified 23 admixtures correctly as such, which gives our model a sensitivity of 88%. The results are summarized as a confusion matrix in Fig. 3.

We want to highlight the difficulties of discriminating Alphonso samples from these self-mixed admixtures containing 35% of non-Alphonso samples. For this, we again display levels of niacin and histidine as a scatterplot in Fig. 4a. Before, we had shown (see Fig. 2a) that this feature pair easily allows for differentiation of Alphonso samples and pure non-Alphonso samples. However, from Fig. 4a, it is apparent that the usefulness of this feature pair for differentiating Alphonso samples and self-mixed admixtures is limited. To assess feature importance for detecting admixtures (at an admixture proportion of 35%), we used information gain just as before. Trigonelline now shows the highest discriminative power ($IG = 0.326$) even though its information gain is significantly diminished compared to before. Interestingly, alanine is the only metabolite whose information gain has not significantly diminished ($IG = 0.224$). This highlights the importance of alanine levels for our model (see also Fig. 4b). Histidine ($IG = 0.210$) and choline ($IG = 0.162$) still show relevant discriminative power. Four metabolites however, namely, niacin, formic acid, adenosine, and shikimic acid, no longer show relevant discriminative power (all $IG = 0$).

For a more thorough understanding of the capabilities of our model, we created virtual admixtures by calculating linear combinations of previously measured samples at different admixture proportions (5, 20, 35, 50, 65, 80, and 95%). For each admixture proportion, 40 samples were

Figure 4 **a** Scatterplot of niacin and histidine levels. Differentiating Alphonso samples and self-mixed admixtures by this feature pair is difficult even though it allowed for easy differentiation of Alphonso samples and pure non-Alphonso samples (see Fig. 2a). An Alphonso sample from the training set that is an outlier was omitted for clarity reasons. **b** Scatterplot of trigonelline and alanine levels. This feature pair is powerful for differentiating Alphonso samples and self-mixed admixtures.

calculated by random drawing. This was repeated ten times at each admixture proportion. Then, we evaluated these virtual admixtures with our model. Results of this are shown in Fig. 5. Even if only 20% of non-Alphonso man-goes are admixed, our model will still detect admixtures with a sensitivity of about 46%. For an admixture proportion of 50%, admixtures will be detected with a sensitivity of 100%.

Fig. 5 Simulated sensitivities of the model at different admixture proportions. Points represent the mean of repeated simulations that are performed to reduce the variance introduced by random drawing. The error bars represent one standard deviation. A logistic function is fit to the data. As a result, even if only 20% of non-Alphonso mangoes are admixed, our model will still detect these admixtures with a sensitivity of 46%

Discussion

For our model, we only retrieve information of eight metabolites. In comparison, most other studies investigating food authenticity by NMR spectroscopy use spectral binning to retrieve a large number of features (Esslinger et al., 2014). However, spectral binning potentially deteriorates information because it is not always possible to dissect overlapping signals. For mango samples of the Alphonso cultivar, this is especially true for signals of histidine, adenosine, and formic acid that overlap with a broad background signal originating from an oligomeric species. Additionally, the model developed in this study relies on a density-based approach. It is well-known that distance- and density-based methods do not perform well on high-dimensional data, especially if they are fed with a relevant number of features that show significant variance but little discriminative power (Aggarwal and Yu, 2001). The number of features used in our study is relatively small, but each feature has significant discriminative potential as assessed by calculating information gains.

Our model for admixture detection is based on the LOF algorithm, which is an anomaly detection algorithm. In comparison, most other studies investigating food authenticity by NMR spectroscopy use supervised machine-learning classifiers. In our opinion, anomaly detection algorithms are better suited to detect admixtures because falsifiers can be expected to create innovative fraud schemes. These cannot fully be covered in the training data fed to supervised

machine-learning classifiers and will diminish their performance in real-world applications. For example, in this study, admixtures were generated by blending six non-Alphonso cultivars to Alphonso samples. However, hundreds of mango cultivars are known. It is simply unfeasible to create training data that covers all of these cultivars. Results shown in this study demonstrate the capabilities of the anomaly detection approach. Admixture detection is highly challenging if the admixed component does not introduce compounds that are completely absent in authentic products. This applies to our work, and we therefore think that our model's ability to detect admixtures of 35% non-Alphonso cultivars to Alphonso samples with a sensitivity of 88% is a very good result.

We obtained sensitivities of our model at different admixture proportions by simulation. For this, we calculated linear combinations of previously measured samples. For an admixture proportion of 35% non-Alphonso cultivars, both experimental and simulated sensitivities were retrieved. Good agreement between experimental and simulated sensitivities was observed (88 vs. 92%). This demonstrates the validity of the simulation approach.

Detection limits of our method might appear relatively high compared to DNA-based methods. For fruit-based products however, DNA-based methods exhibit significantly higher detection limits compared to meat-based products. For example, Scott et al. detected addition of grapefruit to orange juice with a relatively high limit of detection of only 10% (Scott and Knight, 2009). Additionally, typical (processing) conditions for fruit-based products such as low pH and high temperatures (commonly encountered during sterilization) may significantly hamper the success of DNA-based methods (Bauer et al., 2003).

We look forward to see how our model will perform on admixtures of other cultivars not analyzed in this work in the future. This might obviously reveal cultivars that are more similar to the Alphonso cultivar than those investigated in this study. However, we expect our model to handle unexpected samples more robustly than models based on supervised machine-learning classifiers. Anomaly detection algorithms are designed with explicit consideration of irregular occurrences, and therefore, the performance of our model is less likely to diminish. Furthermore, it cannot be expected that the small number of Alphonso samples analyzed in this study can be representative for the entirety of Alphonso mangoes grown. However, in our model, we selected a cut-off value for outlier detection that includes a margin and expands significantly beyond those Alphonso samples observed in the training set. Therefore, we expect that the specificity of our model will not greatly diminish when analyzing future Alphonso samples. In fact, we anticipate that by measuring a greater sample size of Alphonso samples in the future, it will be possible to reduce this margin and thereby improving the performance of our method.

Conclusion

We demonstrate that ^1H -NMR spectroscopy allows for differentiation of the Alphonso cultivar from other mango cultivars, e.g., by elevated niacin, trigonelline, and histidine but relatively low alanine levels. Furthermore, detecting admixture of non-Alphonso cultivars to Alphonso purée is possible using appropriate multivariate data analysis. On a test set, the model developed in this work was able to uncover admixtures consisting of 35% non-Alphonso and 65% Alphonso mango purée with a sensitivity of 88%. At the same time, Alphonso samples were verified with a good specificity of 86%. Detecting such admixtures is highly relevant. A relevant share of the industrial production of Alphonso mangoes is traded as purée. Given the price gap between Alphonso and other mango cultivars, we are convinced that admixture to Alphonso mango purées in industrial production is likely. The model for admixture detection developed in this work shows promising results. However, larger sample sizes will be necessary to validate the performance of our model.

Methodologically, our work demonstrates that calculating second-order lag-one differences of segments of NMR spectra allows for effective baseline flattening. As a result, NMR signals can be easily fitted in cases that would be difficult otherwise. Furthermore, our work highlights the usefulness of anomaly detection algorithms for admixture detection. Algorithms such as the local outlier factor allow that authentic samples become the focus of admixture detection during training. In contrast, for supervised classification algorithms, it is necessary to spend significant resources on analyzing non-authentic samples.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s12161-021-02116-y>.

Acknowledgements We kindly thank SGF International e.V. for providing sample material. Furthermore, we thank Philipp Ewald, Florian Birk, Simone Schmidt, and Dr. Fred Siewek (all Chelab Dr. V. Ara GmbH & Co. KG) for their contribution of ideas.

Author Contribution Conceptualization, Claas Strecker & Victor Ara; Methodology, Claas Strecker; Software, Claas Strecker; Formal Analysis, Claas Strecker; Investigation, Claas Strecker; Resources, Victor Ara; Writing – Original Draft, Claas Strecker; Writing – Review & Editing, Claas Strecker & Victor Ara; Supervision, Victor Ara.

Data availability Raw NMR data available upon reasonable request.

Code Availability Code available upon reasonable request.

Declarations

Ethics Approval This article does not contain any studies with human participants or animals performed by any of the authors.

Informed Consent Informed consent not applicable.

Conflict of Interest At the time of submission, Victor Ara was a shareholder at Chelab Dr. V. Ara GmbH & Co. KG. At present, Victor Ara is no longer a shareholder but consults Chelab Dr. V. Ara GmbH & Co. KG. Claas Strecker is employed at Chelab Dr. V. Ara GmbH & Co. KG.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Ackermann SM, Lachenmeier DW, Kuballa T, Schütz B, Spraul M, Bunzel M (2019) NMR-based differentiation of conventionally from organically produced chicken eggs in Germany. *Magn Reson Chem* 57:579–588. <https://doi.org/10.1002/mrc.4838>
- Aggarwal CC, Yu PS (2001) Outlier detection for high dimensional data. *ACM SIGMOD Rec* 30:37–46. <https://doi.org/10.1145/376284.375668>
- Bachmann R, Klockmann S, Haerdter J, Fischer M, Hackl T (2018) ^1H NMR Spectroscopy for determination of the geographical origin of hazelnuts. *J Agric Food Chem* 66:11873–11879. <https://doi.org/10.1021/acs.jafc.8b03724>
- Bachmann R, Shakiba N, Fischer M, Hackl T (2019) Assessment of mixtures by spectral superposition. An approach in the field of metabolomics. *J Proteome Res* 18:2458–2466. <https://doi.org/10.1021/acs.jproteome.8b00985>
- Bandyopadhyay C, Gholap AS (1973) Changes in fatty acids in ripening mango pulp (var Alphonso). *J Agric Food Chem* 21:496–497. <https://doi.org/10.1021/jf60187a027>
- Bauer T, Weller P, Hammes WP, Hertel C (2003) The effect of processing parameters on DNA degradation in food. *Eur Food Res Technol* 217:338–343. <https://doi.org/10.1007/s00217-003-0743-y>
- Bernstein MA, Sýkora S, Peng C, Barba A, Cobas C (2013) Optimization and automation of quantitative NMR data extraction. *Anal Chem* 85:5778–5786. <https://doi.org/10.1021/ac400411q>
- Borchers HW (2021). *pracma: Practical Numerical Math Functions*. R package version 2.3.3. <https://CRAN.R-project.org/package=pracma>. Accessed 17 June 2021
- Breunig MM, Kriegel H-P, Ng RT, Sander J (2000) LOF: identifying density-based local outliers. *ACM SIGMOD Rec* 29:93–104. <https://doi.org/10.1145/335191.335388>
- Chen X, Lu L, Xiong X, Xiong X, Liu Y (2020) Development of a real-time PCR assay for the identification and quantification of bovine ingredient in processed meat products. *Sci Rep* 10:2052. <https://doi.org/10.1038/s41598-020-59010-6>
- Deshpande AB, Anamika K, Jha V, Chidley HG, Oak PS, Kadoo NY, Pujari KH, Giri AP, Gupta VS (2017) Transcriptional transitions in Alphonso mango (*Mangifera indica* L.) during fruit development and ripening explain its distinct aroma and shelf life characteristics. *Sci Rep* 7:8711. <https://doi.org/10.1038/s41598-017-08499-5>

- Duarte IF, Goodfellow BJ, Gil AM, Delgadillo I (2005) Characterization of mango juice by high-resolution NMR, hyphenated NMR, and diffusion-ordered spectroscopy. *Spectrosc Lett* 38:319–342. <https://doi.org/10.1081/SL-200058713>
- Esslinger S, Riedl J, Fahl-Hassek C (2014) Potential and limitations of non-targeted fingerprinting for authentication of food in official control. *Food Res Int* 60:189–204. <https://doi.org/10.1016/j.foodres.2013.10.015>
- Ferrier P, Peterson E, Landes M (2012) Specialty crop access to U.S. markets: a case study of Indian mangoes, ERR-142, U.S. Department of Agriculture, Economic Research Service. <https://doi.org/10.22004/ag.econ.262228>
- Gil AM, Duarte IF, Delgadillo I, Colquhoun IJ, Casascelli F, Humpfer E, Spraul M (2000) Study of the compositional changes of mango during ripening by use of nuclear magnetic resonance spectroscopy. *J Agric Food Chem* 48:1524–1536. <https://doi.org/10.1021/jf9911287>
- Godelmann R, Fang F, Humpfer E, Schütz B, Bansbach M, Schäfer H, Spraul M (2013) Targeted and nontargeted wine analysis by ¹H NMR spectroscopy combined with multivariate statistical analysis. Differentiation of important parameters: grape variety, geographical origin, year of vintage. *J Agric Food Chem* 61:5610–5619. <https://doi.org/10.1021/jf400800d>
- Gowda GAN, Raftery D (2014) Quantitating metabolites in protein precipitated serum using NMR spectroscopy. *Anal Chem* 86:5433–5440
- Hong E, Lee SY, Jeong JY, Park JM, Kim BH, Kwon K, Chun HS (2017) Modern analytical methods for the detection of food fraud and adulteration by food category. *J Sci Food Agric* 97:3877–3896. <https://doi.org/10.1002/jsfa.8364>
- Hu Y, Murray W, Shan Y (2020). Rlof: R parallel implementation of local outlier factor (LOF). R package version 1.1.2. <https://CRAN.R-project.org/package=Rlof>. Accessed 17 June 2021
- Koda M, Furihata K, Wei F, Miyakawa T, Tanokura M (2012) Metabolic discrimination of mango juice from various cultivars by band-selective NMR spectroscopy. *J Agric Food Chem* 60:1158–1166. <https://doi.org/10.1021/jf2041438>
- McGrath TF, Haughey SA, Patterson J, Fahl-Hassek C, Donarski J, Alewijn M, van Ruth S, Elliott CT (2018) What are the scientific challenges in moving from targeted to non-targeted methods for food fraud testing and how can they be addressed? – spectroscopy case study. *Trends Food Sci Technol* 76:38–55. <https://doi.org/10.1016/j.tifs.2018.04.001>
- Mike FC (2021). ggpattern: Geoms with Patterns. <http://github.com/coolbutuseless/ggpattern>. Accessed 17 June 2021
- Oliveri P (2017) Class-modelling in food analytical chemistry: development, sampling, optimisation and validation issues – a tutorial. *Anal Chim Acta* 982:9–19. <https://doi.org/10.1016/j.aca.2017.05.013>
- Pandit SS, Kulkarni RS, Chidley HG, Giri AP, Pujari KH, Köllner TG, Degenhardt J, Gershenzon J, Gupta VS (2009) Changes in volatile composition during fruit development and ripening of ‘Alphonso’ mango. *J Sci Food Agric* 89:2071–2081. <https://doi.org/10.1002/jsfa.3692>
- R Core Team (2020). R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org/>
- Romanski P, Kotthoff L, Schratz P (2021). FSelector: selecting attributes. R package version 0.33. <https://CRAN.R-project.org/package=FSelector>. Accessed 17 June 2021
- Ryu S, Koda M, Miyakawa T, Tanokura M (2017) Quantitation of minor components in mango juice with band-selective excitation NMR spectroscopy. *J Agric Food Chem* 65:9547–9552. <https://doi.org/10.1021/acs.jafc.7b03336>
- Schmitt C, Bastek T, Stelzer A, Schneider T, Fischer M, Hackl T (2020) Detection of peanut adulteration in food samples by nuclear magnetic resonance spectroscopy. *J Agric Food Chem* 68:14364–14373. <https://doi.org/10.1021/acs.jafc.0c01999>
- Schoenberger T, Menges S, Bernstein MA, Pérez M, Seoane F, Sýkora S, Cobas C (2016) Improving the performance of high-precision qNMR measurements by a double integration procedure in practical cases. *Anal Chem* 88:3836–3843. <https://doi.org/10.1021/acs.analchem.5b04911>
- Schölkopf B, Williamson RC, Smola A, Shawe-Taylor J, Platt JC (1999) Support vector method for novelty detection. *Adv Neural Inf Process Syst* 12:582–588. <https://doi.org/10.5555/3009657.3009740>
- Scott M, Knight A (2009) Quantitative PCR analysis for fruit juice authentication using PCR and laboratory-on-a-chip capillary electrophoresis according to the Hardy–Weinberg law. *J Agric Food Chem* 57:4545–4551. <https://doi.org/10.1021/jf9002686>
- Signal Developers (2013). signal: signal processing. URL: <http://r-forge.r-project.org/projects/signal/>. Accessed 17 June 2021
- Spink J, Moyer DC (2011) Defining the public health threat of food fraud. *J Food Sci* 76:R157–R163. <https://doi.org/10.1111/j.1750-3841.2011.02417.x>
- Varadhan R, Borchers HWW, Bechard V (2020) dfoptim: derivative-free optimization. R package version 2020.10–1. <https://CRAN.R-project.org/package=dfoptim>. Accessed 17 June 2021
- Wickham H (2016) ggplot2: elegant graphics for data analysis. Springer-Verlag, New York
- Wickham H (2019) stringr: simple, consistent wrappers for common string operations. R package version 1.4.0. <https://CRAN.R-project.org/package=stringr>. Accessed 17 June 2021
- Wickham H, François R, Henry L, Müller K (2021). dplyr: a grammar of data manipulation. R package version 1.0.4. <https://CRAN.R-project.org/package=dplyr>. Accessed 17 June 2021
- Wider G, Dreier L (2006) Measuring protein concentrations by NMR spectroscopy. *J Am Chem Soc* 128:2571–2576

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.