Current Research in Food Science 4 (2021) 45–52

Contents lists available at ScienceDirect

Current Research in Food Science

journal homepage: www.editorialmanager.com/crfs/

Research Paper

Rapid classification of commercial teas according to their origin and type using elemental content with X-ray fluorescence (XRF) spectroscopy

Cia Min Lim, Manus Carey, Paul N. Williams, Anastasios Koidis *

Institute for Global Food Security, School of Biological Sciences, Queen's University Belfast, UK

ARTICLE INFO

Keywords:
Tea
Chemometrics
Trace element
Metals
Origin
XRF spectroscopy
Classification

ABSTRACT

The authenticity of tea has become more important to the industry while the supply chains become complex. The quality and price of tea produced in different regions varies greatly. Currently, a rapid analytical method for testing the geographical origin of tea is missing. XRF is emerging as a screening technique for mineral and elemental analysis with applications in the traceability of foodstuffs, including tea. This study aims to develop a reliable multivariate classification model using XRF spectroscopy to obtain the mineral content. A total of 75 tea samples from tea producing countries throughout the world were analysed. After variable shortlisting, 18 elements were used to construct the multivariate models. Tea origin was determined by classifying the tea into 5 major geographical regions producing most of the global tea. PCA showed initial clustering in some regions, although the types of teas included in the study (black, green, white, herbal) showed no discrete cluster membership. The prediction power of each classification model developed was determined by using two multivariate classifiers, SIMCA and PLS-DA, against an independent validation set. The average overall correct classification rates of PLS-DA models were between 54-85% while the results of SIMCA models were between 70-84% resolving the poor clustering initially shown by PCA. This study demonstrated the potential of geographical origin of tea prediction using elemental contents of tea. Naturally, the classification can be linked not only to origin but to the type of tea as well.

Practical application: Wholesalers and retailers need a rapid and robust screening tool to confirm the origin and type of tea they sell to consumers. X-Ray fluorescence spectroscopy proved a good technique for achieving this in commercial teas sourced worldwide. Building on multivariate models, broad classification was accomplished both in terms of origin (Asian vs non-Asian) and in tea type with zero sample preparation and low cost of analysis.

1. Introduction

Consumer interest in the origin of food products has drastically increased over the last decade and an increasing number of products have been marketed on the basis of their origin. Food authenticity has become more important, because of numerous global/severe food adulteration or contamination incidents (Luykx and van Ruth, 2008). All these incidents lead to critical economic losses and concerns about human health (Peng et al., 2017). As a result, determination of food authenticity is an important issue in quality control and food safety (Drivelos and Georgiou, 2012).

Tea, produced from the *Camelia sinensis* leaves, is one of the most widely consumed flavoured and functional beverages worldwide due to its refreshing taste and desirable aroma. The chemical composition of tea, which includes carbohydrates, amino acids, proteins, minerals, polyphenols and alkaloids, provides potential health benefits (Karori et al., 2007; Bogdanski et al., 2012) and important physiological properties (Chang et al., 2017). The characteristics, quality and reputation of tea is affected by its geographical origin (He et al., 2012). The aroma and taste of tea is greatly affected by the geographical location of the tea plant and the natural conditions in which it has been grown (Yan et al., 2014). This is because different cultivating areas provide variations in growing conditions for tea. These variations include climate, rainfall, altitudes, soil, fertilizer, microelements and processing procedures (Yan et al., 2014). These are all contributing factors to the chemical composition in tea leaves, which is related to the quality of teas (Heaney et al., 2018).

Plantations of some specific countries have a better reputation for producing high quality tea, and these producers tend to price their products significantly higher than the average (Ye, 2012). However, tea from other provinces produced using a similar process can hardly be

* Corresponding author. Institute for Global Food Security, School of Biological Sciences, Queen's University Belfast, 19 Chlorine Gardens, Belfast, BT9 5DJ, UK.
E-mail address: t.koidis@qub.ac.uk (A. Koidis).

https://doi.org/10.1016/j.crfs.2021.02.002
Received 17 October 2020; Received in revised form 29 January 2021; Accepted 2 February 2021

distinguished by appearance from the higher quality teas (Zhao et al., 2017a). This encourages dishonest producers to fraudulently label their product as coming from one of the higher quality areas in order to gain a higher price (Ye, 2012). The change in consumer behaviour and the progressively increasing consumption of tea has increased public awareness regarding tea quality and health benefits. To fulfil the consumer demand for high quality tea products, it is necessary to use the appropriate analytical tools when analysing tea authenticity, so that a high standard of quality assurance and process control is maintained (Chen et al., 2015). These techniques can aid in the food traceability system to verify and prevent fake products and misinformation in the event of fraud or commercial disputes (He et al., 2015).

Traditionally, experts were able to discern the origin of a particular tea partly by taste and aroma (Wu et al., 2016). However, this is not always trustworthy (Yu et al., 2009). Quality and flavour parameters (e.g., catechins and polyphenols) analysed by conventional chemical methods such as liquid or gas chromatography are not ideal authenticity markers. This is because they are intrinsically affected by storage time and technological processing of tea (Ye, 2012). Vibrational spectroscopy, which is otherwise useful in authenticity studies, and in this case FT-NIR, has been applied in the context of green tea origin (Chen et al., 2009) and black tea geographical traceability (Ren et al., 2013) with promising results. FT-NIR spectroscopy identification, however, is also linked with variable contents of organic components (caffeine, catechins, polyphenols and free amino acids) and, thus, affected by the same fundamental issues. On the other hand, mineral composition has the potential to determine the geographical origin of the tea because it is less subject to change during production and storage and is thus more stable and more informative when compared to other quality/health related compounds such as antioxidant content. However, it is important to note that the mineral composition of tea leaves not only depends on geographical origin, but also on other factors such as type of soil, tea variety, weather conditions and other seasonal changes (Zhao et al., 2017b). Commercial samples integrate all these aspects/factors and hence why a market basket sampling approach was adopted in this study.

One of the analytical techniques to determine mineral/elemental content is X-ray Fluorescence (XRF) spectroscopy. XRF spectroscopy is rapid, accurate and non-destructive, and only requires minimal sample preparation, in contrast with the more accurate and more expensive Inductive Couple Plasma Mass Spectroscopy (ICP-MS) technique. This solid-state analysis can detect elements ranging from sodium to uranium and has been proven as an effective analytical tool in the determination of elemental (mineral) content in food (Brito et al., 2017; Perring and Andrey, 2018). The technique enables the undertaking of a direct multi-element screening of samples over a wide dynamic range. The fluorescence produced is detectable via XRF spectroscopy after bombardment with high energy X-rays (Van Grieken and Markowicz, 2002; Borgese et al., 2015). XRF spectroscopy is effective with liquid and solid samples, and the wavelength of characteristic X-ray lines is not dependent on the chemical or physical state of the element. This is due to there being no electrons involved in chemical bonding during electronic transitions (Ibañez and Cifuentes, 2001).

The are two XRF variants, Wavelength Dispersive X-ray Fluorescence (WDXRF) and Energy Dispersive X-ray Fluorescence (EDXRF). Here we focus on approaches for the EDXRF systems because they are more common and easier to use (Willis et al. 2014), therefore are preferred for high throughput tea testing. As XRF spectroscopy has a multi-parameter output, it can be used in combination with multivariate untargeted classification techniques such as Partial Least Square-Discriminate Analysis (PLS-DA) and Soft Independent modelling of class analogy (SIMCA). These two supervised linear classification techniques are widely used in qualitative fingerprint analysis of foods (Berrueta et al., 2007). Due to different classification principle, these two techniques are a complementary and often used together.

In this regard, a review of the scientific literature in the authentication of tea origin reveals that most of the studies limited their tea sample collection to a small number of regions within a specific country (Yan et al., 2014). There have, to date, been no studies examining tea samples from different sources around the world. In the one study that used XRF spectroscopy to determine the geographical origin of tea (Rajapaksha et al., 2017), the scope of the experimental design and especially the extent of multivariate analysis employed, in terms of different techniques and appropriate validation, were limited and thus, the research question remains.

Therefore, the objective of the current study is to investigate the suitability of XRF spectroscopy to be used as the basis for rapid determination of tea's geographic origin. Coupling XRF data with a robust modelling approach the aim was to develop and validate a wider classification model using modern chemometric techniques and robust validation.

2. Materials & methods

2.1. Sample collection

Several commercial tea samples (n = 75) were collected for this study of which 54 derived from trusted UK wholesalers between 2017 and 2019. Another 21 samples were collected from trusted sources in Singapore, Malaysia and China during the same period. The types of tea collected were: black (n = 14), green (n = 22), oolong (n = 5), blended (n = 14), rooibos (n = 3), and herbs and fruits tea (n = 17). The suppliers were able to confirm the country of origin, or in terms of blends, the originated countries of the particular batch of tea with a high level of confidence using supply chain tracing documents. Other information about the samples such as packaging format, level of caffeine where available, point of purchase was also collected. The samples were assigned to different "classes" which represent different geographical regions according to the model design (see 3.1).

2.2. Elemental analysis using X-ray fluorescence (XRF) spectroscopy

Dried tea leaves were finely ground using a pestle and mortar before being placed inside the XRF sample cups. The procedures of preparing the tea samples for XRF analysis were adapted from Signes-Pastor et al. (2017) and Afroz et al. (2019). Briefly, the tea powders (2.50–3.00 g) were placed inside two 32 mm double open-ended XRF sample cups (Elementec, Maynooth, Ireland) and covered on one side with Prolene thin-film (Chemplex Industries, Florida, USA). The samples were compressed at 200 psi for 25 sec to a depth ca. 4 mm. The depth and weight of the tea samples were recorded. The elements contained in the tea samples were analysed using an EDXRF spectrometer (Rigaku Nex CG bentchtop XRF, Texas, USA). Full experimental setup and acquisition parameters are provided in Supplementary Material I and the EDXRF Application note: agricultural soils & plant materials (Rigaku, 2015). The instrument was calibrated using the Fundamental Parameter method. The measurement of elements was done by helium purging to enhance Mg and P sensitivity. The samples were analysed in 10 batches, with each batch containing at most 8 tea samples and 1 certified reference material (CRM) in order to determine the accuracy and reproducibility of the measurements (Oriental Tobacco Leaves, CTA-OTL-1). Repeatability study (n = 8) was conducted successfully and allowed running of only one XRF analysis per tea sample (Supplementary Information II).

2.3. Descriptive analysis and multivariate analysis

The concentration of each element was automatically calculated from the XRF spectra in addition to the statistical error, the limit of detection and quantification by the Rigaku RPF-SQX ("profile fitting–spectra quant X") software supplied with the XRF spectrometer. These results (24 elements initially) coupled with TPC were transferred to spreadsheet editor and further descriptive analysis was carried out. Descriptive analysis (mean, standard deviation, maximum and minimum of each element)

was carried out using Excel. After excluding some elements (see 3.1), the dataset (18 elements x 75 samples) was imported to the chemometrics software package, SIMCA 15 (Umetrics, Sweden) for the following multivariate analysis.

The exploration of the dataset was performed using Principal Component Analysis (PCA) both for the raw data and after the application of each data pre-treatment methods. The classification analysis to determine the geographic origin of tea was carried out using two supervised multivariate classification techniques: PLS-DA and SIMCA. According to the model design (see results) the tea samples from each class were separated into a calibration set (70% of total samples) and a validation set (30% of total samples) randomly using the INDEX function in Microsoft Excel. This was done four times to ensure each of the samples were present in at least in one validation set. This iteration enables the calculation of the average classification success and the standard deviation and importantly removes the bias from picking a single validation set (Westad and Marini, 2015). Cross validation (venetian blinds) was also performed using the software independently of the technique chosen, as follows: the data was divided into 7 portions. A model was constructed based on the 6/7ths of data left; the excluded 1/7th was then predicted from the model. This was done repeatedly using each 1/7th of the data until all data portions had been predicted. Predicted Residual Sum of Square (PRESS), i.e., the comparison of the predicted data with the original data and the sum of square errors, Q2 (i.e., how well the model predicts new data) and R2 (how well the model fits the data) were computed in SIMCA 15.

Results are expressed as a percentage of sensitivity specificity and correct classification rate. Sensitivity represents how well the model classifies the target samples, while specificity represents the performance of the model in terms of classifying non-target samples.

3. Results and discussion

3.1. Repeatability of the analysis and variable shortlisting

There are 24 elements detected in the CRM and the selected commercial tea sample following XRF analysis. The repeatability was determined in both cases, but more emphasis was given in the results obtained from the tea sample. Overall, the relative standard deviation (RSD) was lower than 25% except for Co, Pb, U, Pd, Ta and Re which exhibited very high RSD due to their very low content and high standard deviation (data not shown). XRF is known to produce quantification errors in low concentration levels (broadly, <3 mg/kg) due to its sensitivity (Markowicz, 2011). The average RSD of all elements was 49.75%. However, the RSD without Co, Pb, U, Ta and Re decreased to 12.44%. The shortlisting on the elements was performed with the following criteria: a) the element obtained mean concentration of all tea samples over 3 mg/kg (to ensure reliability), b) the element exhibits a RSD <25% and c) the recovery of or the particular element in the CRM was in the range of 65%–125%. The recoveries of most of the elements analysed were acceptable (65–110%), with the exception of As, Na, U and Ba. The elements Co and Pd were not detected in the CRM samples. There is neither recommended value nor information value for the concentration of elements Si, Ti, Sn, Ta and Re in the CRM samples, thus, the recovery data for these elements cannot be retrieved. The recovery percentage data of the CRM is available in the Supplemental material II (Table S2). The elements Co, Pd, Ta and Re were excluded due to their high RSD derived from their low concentration. A total of 18 elements (Mg, Al, P, S, Cl, K, Ca, Mn, Fe, Cu, Zn, Br, Rb, Sr, Ni, Si, Ti and Sn), all showing good repeatability and accuracy results, were selected for further descriptive statistical and chemometrics analysis. Of these 18, 10 were used in the study of Han et al. (2014) and 13 were used in the study of Rajapaksha et al. (2017). Ni and Mn, which were the main elements in the discrimination of the tea samples in Han et al. (2014), are included in this study. However, Na and Pb, which had previously proven to be good indicators of the geographical origin of teas, together with Mg, Ca, Ni, Rb and Sr in the work of Zhao et al. (2017b), did not achieve good sensitivity in the XRF analysis, and were therefore excluded in this study.

3.2. Descriptive statistical analysis of elements in tea samples

The tea samples were initially grouped by country of origin, but several model iterations showed this approach was not suitable due to the low number of samples for every country and the unbalanced group model design. Two model designs were adopted for origin classification and they were region-specific rather than country specific: In Design I, tea samples were divided into five classes according to their geographical origin: “East Asia or E Asia” (n = 22), “South Asia or S Asia” (n = 6), “South East Asia or SE Asia” (n = 7), “Africa” (n = 6), “Other” (n = 34). In Design II, the model is further simplified with two only classes (“Asia” and “non-Asia”). We recognise however that this division does not represent the tea plantation regions globally and does not cover the variability of different teas produced within every region.

In Design I (5 geographic regions or 5 class-model), the concentration of the 18 elements is shown for all samples in Table 1. The full elemental table and repeatability is available in the Supplementary Material II (Table S1). In general, the levels of the minerals following the XRF analysis are similar to those from Rajapaksha et al. (2017) and Yemane et al. (2008), whereby Mg, P, S, K and Ca are the primary elements found in tea samples, with K the most abundant. This study confirms a high concentration of K from the tea collected from all regions. The African class has the lowest concentration of element at 9800 mg/kg on average. Other common elements in tea include calcium, magnesium, aluminium, phosphorous and silicon but they are found at a much lower concentration than potassium. Other trends in the data include very low concentrations of Ti in tea from E Asia, Africa and ‘Other’ regions, indirectly suggesting good quality soil in those regions. This is because Ti is not absorbed by the tea plants, so it's a good indicator of direct soil

Table 1
Concentration of 18 elements measured by XRF spectrometer according to class Design I (5 global regions).

Element	East Asia	South Asia	South East Asia	Africa	Others
Mg	2590 ± 797	2370 ± 403	2217 ± 351	2115 ± 460	3020 ± 1385
Al	1553 ± 718	1477 ± 557	1397 ± 346	868 ± 212	1433 ± 678
P	2109 ± 910	1323 ± 138	1451 ± 195	990 ± 513	1424 ± 502
S	2200 ± 573	1822 ± 281	1783 ± 94.1	1206 ± 348	2019 ± 942
Cl	638.7 ± 262	615 ± 228	624.4 ± 258	1405 ± 1110	1660 ± 1821
K	15726 ± 3876	15750 ± 1679	15657 ± 605	9830 ± 6812	15558 ± 4119
Ca	3609 ± 1165	4485 ± 1293	4013 ± 793	2947 ± 980	7169 ± 4556
Mn	845 ± 278	629 ± 403	742 ± 228	576 ± 563	628 ± 586
Fe	150.7 ± 122	119.0 ± 34.7	114.0 ± 39.2	176 ± 51.8	306 ± 193
Cu	13.76 ± 4.7	20.7 ± 2.4	18.6 ± 3.3	10.2 ± 4.64	13.9 ± 4.7
Zn	27.4 ± 9.1	27.7 ± 4.7	23.8 ± 4.3	17.8 ± 5.34	25.7 ± 6.89
Br	3.4 ± 1.3	3.1 ± 1.1	3.1 ± 1.0	16.3 ± 12.7	8.1 ± 6.4
Rb	46.90 ± 28.7	26.58 ± 14.3	62.9 ± 30.4	44.7 ± 42.2	31.1 ± 23.7
Sr	11.2 ± 7.3	18.7 ± 5.9	23.2 ± 20.8	22.9 ± 12.3	41.2 ± 24.7
Ni	5.4 ± 3.5	6.9 ± 2.0	4.8 ± 2.5	3.2 ± 2.8	4.6 ± 2.8
Si	1113 ± 1461	905 ± 628	429 ± 141	923 ± 403	2403 ± 2193
Ti	2.0 ± 5.6	0	0	5.5 ± 6.39	6.9 ± 10.7
Sn	33.6 ± 5.6	31.8 ± 7.2	34.1 ± 2.5	30.7 ± 4.7	31.2 ± 8.2

contamination. Tea from Africa and "Other" had outstanding higher mean values of Cl and Br than Asia. Besides that, SE Asia contains the highest mean value of the element Rb. In addition, tea from E Asia has the highest mean concentration value of elements Al, P, S, and Mn while South Asia has highest mean concentration of K, Cu, Zn and Ni. The lowest mean concentration value of Mg is found in African tea, although the minimum value of Mg is found in tea from the 'Other' region. Some trends with low concentration minerals were also observed indicating their maybe unreliability due to the higher quantification errors.

The descriptive statistics of the elemental contents of different classes showed high standard deviations were observed for a number of elements. This indicates that the concentration of these elements varied greatly within the same region. This variation can be explained either by the different types of tea included in the class, the processing method used, or the ingredients blended with the tea samples. The latter may have affected the elemental intensities of the tea leaves.

3.3. Multivariate analysis: principal component analysis (PCA)

The flowchart shows the summary of the experimental design development for multivariate analysis of this study (Fig. 1). Starting from Design I (five-class model), all tea samples were used to determine the best data pre-treatment procedure. Data pre-treatment is an important step that needs to be undertaken prior to multivariate analysis because it removes undesirable systematic variation from the dataset, thereby enhancing the predictive power of the PLS-DA and SIMCA calibration models (Berrueta et al., 2007; Eriksson et al., 2013). In the relevant literature with elemental analysis and geographical origin approach, Rajapaksha et al. (2017) used half-range and central value transformation for data rescaling, a process that is also recommended in the work of Moreda-Pineiro et al. (2001). However, the multi-element determination method used in Moreda-Pineiro et al. (2001) is ICP-MS/AES), not the XRF employed in this study and the study of Rajapaksha et al. (2017). Here, after testing more than 10 different pre-treatment methods, including UV and Pareto scaling, the Automatic Transform produced the best results (Table S2) and was therefore applied to the data set. On this basis, three outliers (T24, T27 and T46) were excluded from the dataset.

PCA was performed to undertake exploratory data analysis and unsupervised pattern recognition. This approach provides a more comprehensive picture of the XRF-analysed datasets by simplifying and providing a graphical visualisation of the data. Class information is not required to construct the PCA model.

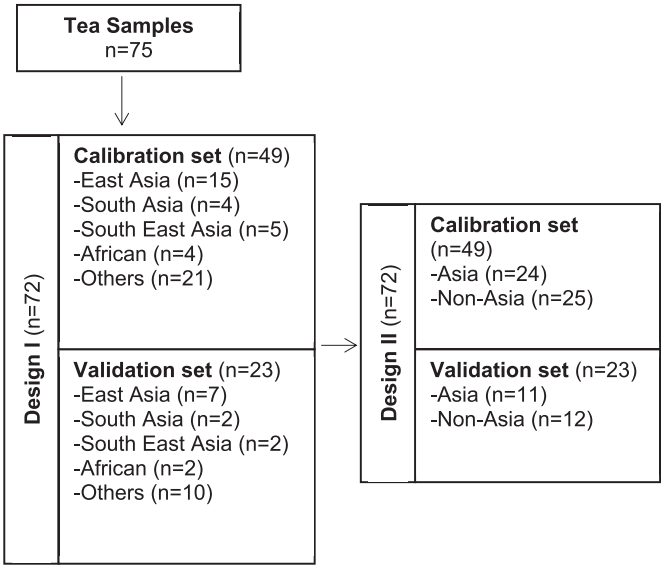

Fig. 1. The experimental design of the study outlining the classes and the split between calibration and validation set for multivariate analysis.

The consequent PCA model produced showed that separation of the 5 geographical regions is not as clear as the tea samples from E Asia, S Asia and SE Asia clustered with some tea samples from African and Others (Fig. 2A). PCA, tea samples from the group "Others" and African classes have a clear separation from other tea samples. The PCA of Design I informed the development of Design II, which merged tea samples in E Asia, S Asia and SE Asia into new class 'Asia' and leaving the rest of the samples in the 'non-Asian' class. This simplification of the model could be beneficial if, for example, a supplier would like to know if a sample is a Kenyan or a Chinese tea. As expected, the PCA model of Design II, with outliers (T24, T27 and T46) excluded, gives a better separation between classes when the initial classes are merged. However, there is some tea samples from the 'non-Asian' class clustered with the Asian tea class (Fig. 2B).

The PCA scores plots (Fig. 2A and B) remain the same, as alterations to the class design do not affect the PCA model. Similarly, the loadings plot for both Design I and Design II PCA models are identical. Interestingly, only two principal components were used to construct the PCA and the first PC accounts for a very high explained variance (>95%). There is not however, one element that is responsible for this. The loadings plots (Supplementary Material II, Fig. S1) reveal that the most discriminative elements within the tea samples are K, Ca and Mg for the first extracted principal component, while Mn, Cl and Si are the most discriminative elements in terms of the second principal component - which is also indicated in the raw data (Table 1).

Insight can be obtained looking at the types of tea used as not all of them were black teas. The African teas (T26, T43 and T64) clustered as a small group are, in fact, all herbal teas, and more specially Rooibos, with very different mineral profile to the rest. These herbal teas are not actually produced from *Camellia sinensis*, but from the plant *Aspalathus linearis* of the *Fabaceae* family. There seems to be three different cluster of the 'Others' category. Samples that are plotted inside the tolerance ellipse (T7, T22, T23, T28, T30, T31, T37, T48, T51, T62 and T63) on the bottom half are actually herbal teas, which only contain tea leaves in varied concentrations blended with other ingredients such as lemongrass, ginger, orange peel, blackberry leaves, fruits pulp, anise, fennel and camomile. The cluster (T24, T27, T36, T46, T49 and T53) that appears as outlier are teas blended with either spearmint or peppermint. Margui and Voutchkov (2018) found that the concentration of Mn and Rb is significantly higher in black tea than mint tea, while the concentration of Ca and Sr is higher in mint tea. The findings from this study concur with those evidenced by Margui and Voutchkov (2018). The mean concentration of Mn, Rb, Ca and Sr of T24, T27, T36, T46, T49 and T53 are 60.28 mg/kg, 11.865 mg/kg, 12.080 mg/kg and 67.75 mg/kg respectively. The mean concentration of Mn, Rb, Ca and Sr of all tea samples are 698.65 mg/kg, 39.45 mg/kg, 5277.47 mg/kg and 27.49 mg/kg respectively. The concentration of Mn and Rb are approximately 10 and 3 times lower respectively in mint tea samples, while the concentration of Ca and Sr are twice as high when compared to all tea samples. The concentration of Mn, Rb, Ca and Sr of T27 was the lowest among all of the mint tea samples. This may explain why T27 sits to the left of the graph while the other mint tea samples are clustered towards the bottom right. According to the loadings plot, precisely these four elements are discriminative in PCA models. The last cluster is in the centre of the plot at it is mostly black tea. On this basis, it appears that classification can be linked not only to the area of origin but to the type of tea as well. In fact, the herbal teas can be clearly defined from the rest because they have so distinctive mineral profile.

3.4. Multivariate analysis: classification using PLS-DA and SIMCA

Prior to the supervised multivariate classification, all the samples were divided into calibration set (70%) and validation set (30%). The purpose of the calibration set is to create the model while the validation

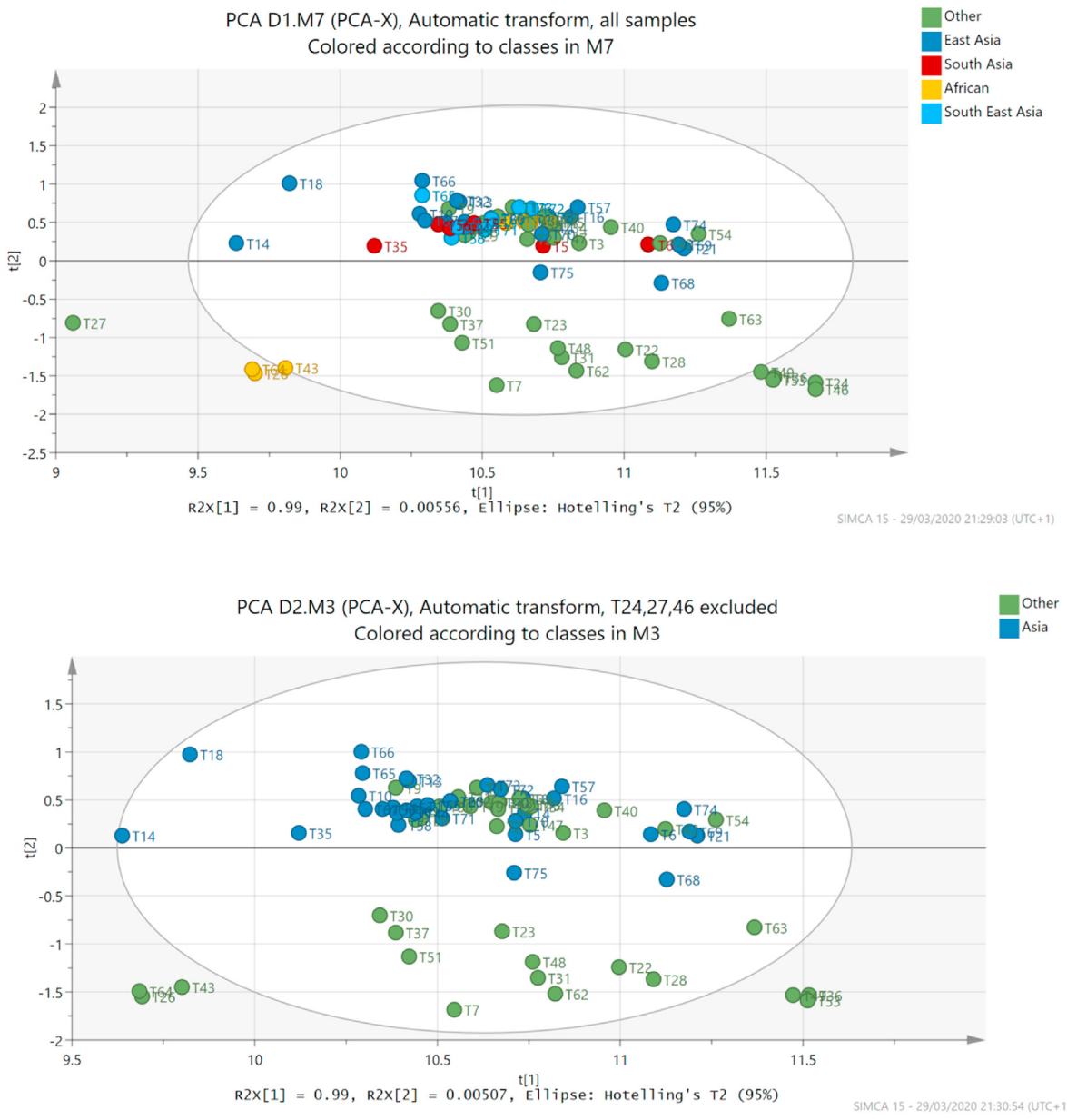

Fig. 2. PCA model of A) Design I (East Asia, South Asia, South East Asia, African and Others) using all tea samples (n = 75); Design II (Asia and Others) using tea samples without outliers (n = 72).

set was used to test the predicted power of the model and this approach has been followed extensively in the literature (Ren et al., 2013). For the classification analysis, two different supervising learning algorithms (Soft Independent modelling of class analogy, SIMCA, and Partial Least Square-Discriminate Analysis, PLS-DA) were used to calibrate the model for 70% of the total tea samples as per the experimental design (Fig. 1). These two pattern recognition methods have different ways of approaching classification: the former method focuses on individual class modelling while the latter method is oriented towards discriminating between the classes in one model that contains several classes (Berrueta et al., 2007).

Cross validation of the calibration set with 1/7th of the data and 7 iterations shows very good model characteristics (R2X, Q2) for the SIMCA models in both Designs I and II (Table 2). More specifically, R2 is the percent of variation of the training set Y and is limited to 1. Q2 indicates the performance of the models in the prediction of new data. The higher Q2, the better the prediction power of the model. The R2X and Q2 of the SIMCA models appeared is very high (value close to 1). On the other hand, the predictability (Q2) of PLS-DA models in all cases (Designs I and II) was very low, which are significantly lower than 0.5 (threshold for acceptable models). The Design I/PLS-DA model achieved the lowest Q2 results (0.134) among all models. This is due to both the data and the

Table 2
Performance parameters of PLS-DA and SIMCA classification models of each design using 18 elemental content of all samples (n = 72).

	Classifier	Class	R2X (cum)	Q2 (cum)
Design I	*SIMCA*	E Asia	0.999	0.996
		S Asia	1.000	0.995
		SE Asia	1.000	0.999
		African	0.997	0.995
		Other	0.997	0.994
	PLS-DA	All Classes	0.998	0.134
Design II	*SIMCA*	Asia	0.996	0.994
		Non-Asian	0.997	0.993
	PLS-DA	All Classes	0.997	0.397

number of classes present; models with higher number of classes will normally exhibit lower prediction power.

The proper test for the model's performance, however, is with a separate prediction set. The classification results of the prediction set (30% of the total samples) for every class is shown in Table 3. Note that this process was repeated four times and four different iterations were performed (see Materials and Methods). Among the two different class designs tested, Design II (two-class model) achieved the highest overall correct classification rate in both PLS-DA and SIMCA models compared to Design I (five-class model), as expected. When comparing PLSDA and SIMCA results it must be noted that the classification algorithm is different, and the results are complimentary rather than mutually exclusive. SIMCA produces independent models for each class (here different types of teas) with PCA and uses Euclidean distance to measure the proximity of the unknown samples to the centre of the class. In contrast, PLS-DA algorithm performs a PLS regression with the Y-variable generated for the class type) mapped into a linear space. The resulting reduced orthogonal space is generated while preserving the maximum linear correlation between the variables and the class type (Reinholds et al., 2015). The Discriminant (DA) part of the analysis is rather straight forward as the regression numbers are tracked back into class memberships as part of the class allocation.

In Design I, comparing the correct classification rate of each class design, it is clear that the SIMCA model is significantly more effective than PLS-DA in predicting the geographical origin of tea samples using elemental data. The small or unbalanced class size affects the models. This was observed with the S Asia class in SIMCA and with three origin classes (S Asia, SE Asia, and Africa) in PLS-DA, explains the 0% correct classification rate. In addition, using the PCA model in Design I it was identified that the African class contains three outliers (rooibos teas). This is known to affect the calibration of models and can lead to inaccurate predictions in validation (Dazykowski et al., 2007). The class 'Others', referring to other teas originated from other than the regions indicated, showed varied classification success. This class in the larger class of the design and such errors can be expected due to unbalances in the overall class membership.

On the other hand, Design II achieved the highest average correct classification rate for both SIMCA (84%) and PLS-DA (85%). In other words, the mineral content of the teas can predict if a sample is Asian or non-Asian tea. One of the explanations for this result is that the total number of classes (i.e., 2) is low, which reduces the likelihood of misclassification.

It is important to be mindful of overfitting. Overfitting takes place when the model loses it generalisation as a result of learning the characteristics of the data. In overfitting, the noise is modelled as well as the unique information (Berrueta et al., 2007). Identifying the relevant variables can prevent this from occurring. The number of variables (18 minerals) selected is considered appropriate for Design I as the total variables used for further analysis exceeds (n-g)/3, where n is the number of samples and g is the number of classes (Defernez and Kemsley, 1997). In Design II however, with only two classes, the variables are higher than (n-g)/3 and as a result, Design II is more prone to overfitting and to prevent than perhaps fewer variable must be used.

Despite that, the results obtained are rather good, taking into consideration that the dataset contains large variability, with the different types of teas (both green and black, and oolong teas), different processing (e.g., microbial fermentation of some traditional teas vs standard oxidation), presence of blends with herbs and fruits and, of course, different origins. Other studies with the same or other methods manage to achieve better classification rate (reaching as high as 94.3% in the case of NIR spectroscopy) controlling carefully the regions selected and limiting the type of tea analysed (Ren et al., 2013).

As discussed earlier, the identification of minerals could be a better method for discriminating geographical origin, as there is less variability between samples compared to that of organic constituents in tea samples. This study deals with realistic dataset representative of the global consumption of teas. There is no doubt that the presence of some herbal and fruit tea samples (n = 20) in the dataset, in addition to tea leaves blended with other ingredients, especially in the 'Others' class, is skewing the data. This reiterates that the classification results are related to both the geographical origin and the type and variety of tea.

Table 3
PLS-DA and SIMCA model average performance on 4 different validation sets for each design.

Model design & Classifiers		Class	Sensitivity	Specificity
Design I	*SIMCA*	E Asia	71.4 ± 11.7	96.8 ± 3.6
		S Asia	25.0 ± 18.9	100.0
		SE Asia	75.0 ± 28.9	86.9 ± 9.8
		African	62.5 ± 25.0	96.4 ± 2.4
		Other	80.0 ± 18.3	78.8 ± 14.6
		Overall	**70.6 ± 9.6**	**89.1 ± 7.7**
	PLS-DA	E Asia	82.1 ± 13.7	59.4 ± 13.0
		S Asia	0	100.0
		SE Asia	0	100.0
		African	0	100.0
		Other	67.5 ± 23.6	69.2 ± 10.8
		Overall	**54.4 ± 13.0**	**83.1 ± 5.8**
Design II	*SIMCA*	Asia	77.3 ± 18.9	89.6 ± 12.5
		Non-Asian	89.6 ± 12.5	77.3 ± 18.9
		Overall	**83.7 ± 10.9**	**83.7 ± 10.9**
	PLS-DA	Asia	88.6 ± 8.7	81.2 ± 7.9
		Non-Asian	81.2 ± 8.0	88.6 ± 8.7
		Overall	**84.7 ± 8.1**	**84.7 ± 8.1**

The XRF method has several advantages. The measurement is non-destructive and requires minimal sample preparation. The contamination risks are low, and results can be obtained within 20 minutes. However, a key limitation of XRF is the method's partial sensitivity in terms of large mass elements. This is due to the high scattering of X-rays, which can lead to increased background intensities. Sodium and lead, which can be important elements in the prediction of geographical origin with other techniques, could not be measured accurately due to the nature of XRF analysis. Na is a light element and XRF quantification is inherently problematic considering the levels found in tea. Similarly, Pb levels detected could be due to anthropogenic contamination. In both cases, the uncertainty of the measurement would not make these two elements good markers for traceability. Nevertheless, XRF is an analytical method that requires frequent validation in terms of precision, accuracy, sensitivity, specificity, uncertainty, and robustness (Berrueta et al., 2007).

Other limitations are related to the sample set used in this experiment. The sample size employed in this study was relatively low, despite efforts to procure representative of tea products from around the world. This was particularly the case in S Asian, SE Asian and African samples. Although much larger samples size (>300) would produce more robust multivariate models, a common issue with studies that use untargeted multivariate analysis, it is important to perform initial investigations like this current one to identify the broad trends. In addition, the tea samples used here, were of different varieties, including green tea, black tea, oolong tea, herbal/fruit tea. Apart from adding both variability and noise to the dataset, some of the herbal tea samples did not contain or contained a trace amount of *Camellia sinensis*. The additional ingredients included in herbal and fruit teas may conceal the effect of characteristic elements of the samples shifting the discrimination from an origin based to a type-of-tea based approach. It is therefore essential to remove them from the dataset when possible (Daszykowski et al., 2007).

4. Conclusion

The rapid classification of tea products according to geographical origin is crucial in the process of quality control in various points along the tea supply chain. In this study, XRF determination was based on a suite of 18 elements in tea with an average limit of detection of above 3 mg/kg to give reliable results. The method is simple, rapid and non-

destructive and can be classified as a screening method. Two different modelling approaches have been followed to broadly discriminate between tea producing regions. The multivariate models produced with either classifier have shown interesting results (>80% correct classification rate in most of the cases). The different type of teas within the different groups has affected the classification, skewing the results, but it is known that a large, varied, real-world dataset adds the necessary model variability that is important in the development of multivariate classification models. Although limited by some factors which have been identified (limited sample size, unbalanced group of origins and types in the dataset), this study shows that the classification of tea samples according to their geographical origin is feasible through the use of the elemental contents measured by XRF and the application of multivariate analysis. The recommendations mentioned previously could assist in improving the overall performance of this method.

CRediT authorship contribution statement

Cia Min Lim: Investigation, Writing - original draft, Writing - review & editing, Formal analysis, Visualization. **Manus Carey:** Validation, Writing - review & editing. **Paul N. Williams:** Validation, Writing - review & editing, Visualization. **Anastasios Koidis:** Conceptualization, Methodology, Supervision, Writing - review & editing, Visualization, Formal analysis, Resources.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Note: Lower Limit of Detection (LLD) is not provided here. LLD depends on many factors including measurement time and the overall composition of the sample (for typical values see Rigaku, 2015).

Acknowledgements

the authors would like to warmly thank Professor Andrew Meharg for providing access to the XRF instrument within the Institute for Global Food Security, QUB.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.crfs.2021.02.002.

Author contributions

LCM contributed to tea sample procurement overseas, performed all the XRF data acquisition, and initial statistical analysis. TK had the overall idea about the research, contributed to sample acquisition, and data interpretation and performed extra statistical work. Both authors equally wrote the manuscript and approved the final version.

References

Afroz, H., Su, S., Carey, M., Meharg, A.A., Meharg, C., 2019. Inhibition of microbial methylation via arsM in the rhizosphere: arsenic speciation in the soil to plant continuum. Environ. Sci. Technol. 53 (7), 3451–3463.

Berrueta, L.A., Alonso-Salces, R.M., Heberger, K., 2007. Supervised pattern recognition in food analysis. J. Chromatogr. A 1158, 196–214.

Bogdanski, P., Suliburska, J., Szulinska, M., Stepien, M., Pupek-Musialik, D., Jablecka, A., 2012. Green tea extract reduces blood pressure, inflammatory biomarkers, and oxidative stress and improves parameters associated with insulin resistance in obese, hypertensive patients. Nutr. Res. 32 (6), 421–427.

Borgese, L., Bilo, F., Dalipi, R., Bontempi, E., Depero, L.E., 2015. Total reflection X-ray fluorescence as a tool for food screening. Spectrochim. Acta, Part B 113, 1–15.

Brito, G.B., Teixeira, L.S.G., Korn, M.G.A., 2017. Direct analysis of marine macroalgae for determination of macro minerals by energy dispersive X-ray fluorescence. Microchem. J. 134, 35–40.

Chang, C.W., Wang, S.H., Jan, M.Y., Wang, W.K., 2017. Effect of black tea consumption on radial blood pulse spectrum and cognitive health. Compl. Ther. Med. 31, 1–7.

Chen, Q., Zhao, J., Lin, H., 2009. Study on discrimination of Roast green tea (Camellia sinensis L.) according to geographical origin by FT-NIR spectroscopy and supervised pattern recognition'. Spectrochim. Acta Mol. Biomol. Spectrosc. 72 (4), 845–850.

Chen, Q., Zhang, D., Pan, W., Ouyang, Q., Li, H., Urmila, K., Zhao, J., 2015. Recent developments of green analytical techniques in analysis of tea's quality and nutrition. Trends Food Sci. Technol. 43 (1), 63–82.

Daszykowski, M., Kaczmarek, K., Heyden, Y.V., Walczak, B., 2007. Robust statistics in data analysis- A review: basic concepts. Chemometr. Intell. Lab. Syst. 85 (2), 203–219.

Defernez, M., Kemsley, E.K., 1997. The use and misuse of chemometrics for treating classification problems. Trac. Trends Anal. Chem. 16 (4), 216–221.

Drivelos, S.A., Georgiou, C.A., 2012. Multi-element and multi-isotope-ratio analysis to determine the geographical origin of foods in the European Union. Trends Anal. Chem. 40, 38–51.

Eriksson, L., Byrne, T., Johansson, E., Trygg, J., Vikstrom, C., 2013. Multi- and Megavariate Data Analysis Basic Principles and Applications. Umetrics Academy, pp. 255–259.

Han, Q., Mihara, S., Hashimoto, K., Fujino, T., 2014. Optimization of tea sample preparation methods for ICP-MS and application to verification of Chinese tea authenticity. Food Sci. Technol. Res. 20 (6), 1109–1119.

He, W., Zhou, J., Cheng, H., Wang, L., Wei, K., Wang, W., Li, X., 2012. Validation of origins of tea samples using partial least squares analysis and Euclidean distance method with near-infrared spectroscopy data. Spectrochim. Acta Mol. Biomol. Spectrosc. 86, 399–404.

He, X., Li, J., Zhao, W., Liu, R., Zhang, L., Kong, X., 2015. Chemical fingerprint analysis for quality control and identification of Ziyang green tea by HPLC. Food Chem. 171, 405–411.

Heaney, S., Koidis, A., Morin, J.M., 2018. Tea and flavoured tea. In: Morin, J.F., Lees, M. (Eds.), Handbook of Food Authenticity: A Guide to Food Authenticity Issues and Analytical Solutions. Elsevier, pp. 315–331. ISBN print version 978-2-9566303-0-2 ; ISBN electronic version 978-2-9566303-1-9. https://doi.org/10.32741/fihb.

Ibañez, E., Cifuentes, A., 2001. New analytical techniques in food science. Crit. Rev. Food Sci. Nutr. 41 (6), 413–450.

Karori, S.M., Wachira, F.N., Wanyoko, J.K., Ngure, R.M., 2007. Antioxidant capacity of different types of tea products. Afr. J. Biotechnol. 6 (19), 2287–2296.

Luykx, D.M.A.M., van Ruth, S.M., 2008. An overview of analytical methods for determining the geographical origin of food products. Food Chem. 107 (2), 897–911.

Margui, E., Voutchkov, M., 2018. Multielement analysis of tea and mint infusions by total reflection X-ray fluorescence spectrometry. Food Anal. Methods 11, 282–291.

Markowicz, A., 2011. An overview of quantification methods in energy-dispersive X-ray fluorescence analysis. Pramana 76 (2), 321–329.

Moreda-Pineiro, A., Marcos, A., Fisher, A., Hill, S.J., 2001. Evaluation of the effect of data pre-treatment procedures on classical pattern recognition and principal components analysis: a case study for the geographical classification of tea. J. Environ. Monit. 3, 352–360.

Peng, G.J., Chang, M.H., Fang, M., Liao, C.D., Tsai, C.F., Tseng, S.H., Kao, Y.M., Chou, H.K., Cheng, H.F., 2017. Incidents of major food adulteration in Taiwan between 2011 and 2015. Food Contr. 72, 145–152.

Perring, L., Andrey, D., 2018. Multi-elemental ED-XRF determination in dehydrated bouillon and sauce base products. Food Anal. Methods 11, 148–160.

Rajapaksha, D., Waduge, V., Padilla-Alvarez, R., Kalpage, M., Rathnayake, R.M.N.P., Migliori, A., Frew, R., Abeysinghe, S., Abrahim, A., Amarakoon, T., 2017. XRF to support food traceability studies: classification of Sri Lankan tea based on their region of origin. X Ray Spectrom. 46 (4), 220–224.

Reinholds, I., Bartkevics, V., Silvis, I., van Ruth, S., Esslinger, S., 2015. Analytical techniques combined with chemometrics for authentication and determination of contaminants in condiments: a review. J. Food Compos. Anal. 44, 56–72.

Ren, G., Wang, S., Ning, J., Xu, R., Wang, Y., Xing, Z., Wan, X., Zhang, Z., 2013. Quantitative analysis and geographical traceability of black tea using Fourier transform near-infrared spectroscopy (FT-NIRS). Food Res. Int. 53, 822–826.

Rigaku, 2015. EDXRF Application note. Agric. Soils Plant Mater. (1385) https://www.rigaku.com/press/wed-04292015-1200/216839527 accessed 28/1/2021.

Signes-Pastor, A.J., Carey, M., Meharg, A.A., 2017. Inorganic arsenic removal in rice bran by percolating cooking water. Food Chem. 234, 76–80.

Van Grieken, R.E., Markowicz, A.A., 2002. Handbook of X-Ray Spectrometry, second ed., vol. 29. Marcel Dekker Inc, New York, ISBN 0-8247-0600-5.

Westad, F., Marini, F., 2015. Validation of chemometric models – a tutorial. Anal. Chim. Acta 893, 14–24.

Willis, J., Feather, C., Turner, K., 2014. Guidelines for XRF Analysis: Setting up Programmes for WDXRF and EDXRF. James Willis Consultants, Cape Town, South Africa.

Wu, Y., Lv, S., Wang, C., Gao, X., Li, J., Meng, Q., 2016. Comparative analysis of volatiles difference of Yunnan sun-dried Pu-erh green tea from different tea mountains: Jingmai and Wuliang mountain by chemical fingerprint similarity combined with principal component analysis and cluster analysis. Chem. Cent. J. 10.

Yan, S.M., Liu, J.P., Xu, L., Fu, X.S., Cui, H.F., Yun, Z.Y., Yu, X.P., Ye, Z.H., 2014. Rapid discrimination of the geographical origins of an oolong tea (*Anxi-Tieguanyin*) by near-infrared spectroscopy and partial least squares discriminant analysis. J. Anal. Methods Chem. https://doi.org/10.1155/2014/704971, 2014.

Ye, N.S., 2012. A mini review of analytical methods for the geographical origin analysis of teas (*Camellia sinensis*). Crit. Rev. Food Sci. Nutr. 52 (9), 775–780.

Yemane, M., Chandravanshi, B.S., Wondimu, T., 2008. Levels of essential and non-essential metals in leaves of the tea plant (*Camellia sinensis L.*) and soil of Wushwush farms, Ethiopia. Food Chem. 107, 1236–1243.

Yu, H., Wang, Y., Wang, J., 2009. Identification of tea storage times by linear discrimination analysis and back-propagation neural network techniques based on the eigenvalues of principal components analysis of E-nose sensor signals. Sensors 9, 8073–8082.

Zhao, H., Zhang, S., Zhang, Z., 2017a. Relationship between multi-element composition in tea leaves and in provenance soils for geographical traceability. Food Contr. 76, 82–87.

Zhao, H., Yu, C., Li, M., 2017b. Effects of geographical origin, variety, season and their interactions on minerals in tea for traceability. J. Food Compos. Anal. 63, 15–20.