Trends in Food Science & Technology 90 (2019) 187–193

Contents lists available at ScienceDirect

Trends in Food Science & Technology

journal homepage: www.elsevier.com/locate/tifs

Commentary

Sampling guidelines for building and curating food authenticity databases

James Donarski[a,*], Federica Camin[b], Carsten Fauhl-Hassek[c], Rob Posey[d], Mike Sudnik[e]

[a] Fera Science Ltd, York, UK
[b] Fondazione Edmund Mach (FEM), San Michele all'Adige, Italy
[c] German Federal Institute for Risk Assessment (BfR), Berlin, Germany
[d] Food Forensics, Norwich, UK
[e] Elementar UK Ltd, Stockport, UK

ABSTRACT

Background: Food fraud is a global issue often detected through the use of analytical testing. Analysis of suspect foodstuffs and comparison of their results to those contained within a food authenticity database is a typical approach. This scientific opinion was commissioned as part of the FoodIntegrity EU project to provide guidance for the creation of these food authenticity databases.

This opinion paper provides what the authors believe are the most important considerations which must be addressed, when creating a food authenticity database. Specifically, the areas of database scope, analytical methodology, sampling, collection and storage of data, validation and curation are discussed.

1. Introduction

This scientific opinion was commissioned as part of the FoodIntegrity EU project to provide guidance for the creation of food authenticity databases.

A food authenticity database is an organised collection of data, analysed with established protocols acquired from a representative number of authentic samples, with the purpose of defining the natural variability of some particular defined property of a foodstuff. This natural variability is taken as a reference and for comparison, when analysing tested samples, to tackle food fraud such as mis-labelling, product extension and adulteration. Given the ultimate aim of such databases, and the implications if a tested food has shown not to conform to a database, it is imperative that specific areas are addressed before, during and after the creation of such a database. These include: definition of the scope of a database; collection of representative, authentic reference materials; sample preparation; data acquisition; validation; database storage/external access; and ensuring collated data remain valid.

Analytical methods for authentication are classified into different types: a) analysis of marker compounds not naturally occurring in the foodstuffs that are characteristic of a particular adulteration (e.g. melamine and other compounds); b) targeted analysis of analytes/markers that are naturally occurring in the foodstuff and comparison of these values to reference data (e.g. the concentration of methylglyoxal for specific active manuka honeys); and c) fingerprinting techniques that simultaneously measure a range of analytes/markers and comparison of these profiles to reference data (e.g. ^{1}H NMR analysis of high value spices). In the case of 'type a' methods, threshold limits are often defined and can be included in specification rules or in regulations In the case of 'type b' and 'c' methods, databases for authentication are of importance.

In the remainder of this manuscript we will refer to 'type b' as 'targeted analysis databases' and 'type c' as 'non-targeted analysis databases'.

Targeted analysis databases are the most established in food authenticity and are used to control several types of food fraud. The methods used are more easily transferred between laboratories than untargeted methods, and this enables more widespread use. To ensure comparability of data, defined robust methods should be used, and it is recommended that the competence of the laboratories using the database, when either providing reference values or using the database to challenge a suspect sample, is verified through the participation in appropriate proficiency testing schemes.

An example of a targeted analysis database is one that contains stable isotope ratios for the purposes of the verifying the geographical origin of food (several examples were discussed in detail by Camin et al., 2017 (Camin et al., 2017)). The most prominent database containing stable isotope ratio data is the so called 'EU-Wine databank' established and operated according to "Commission Regulation (EC) No 555/2008 of 27 June 2008 laying down detailed rules for implementing Council Regulation (EC) No 479/2008 on the common organisation of the market in wine as regards support programmes, trade with third countries, production potential and on controls in the wine sector". In

* Corresponding author.
E-mail address: james.donarski@fera.co.uk (J. Donarski).

https://doi.org/10.1016/j.tifs.2019.02.019
Received 3 July 2018; Received in revised form 13 December 2018; Accepted 6 February 2019
Available online 14 February 2019

that database, isotopic ratios of reference wines from the wine growing countries in Europe are compiled, which have been generated and are up-dated every year by a network of accredited control laboratories using fully validated official methods. The databank wines samples are micro-vinificated according to a harmonised procedure starting from grapes under official control and are considered to be 100% authentic. Due to the high validity of the data, particularly the guaranteed authenticity of the samples, the official EU-Wine databank is considered to be of outstanding quality.

Although its analysed samples are of assured authenticity, the EU-Wine databank relies on experimental wines (micro-vinification) and is functional only for the analysis and appreciation of the stable isotope ratios for which it was developed. The use of such samples is limited, as differences in the composition of these experimental wines and corresponding commercial ones have been reported (Smeyers-Verbeke et al., 2009). One of the main criticisms of authenticity testing in general is that the reference data does not cover the natural and technological diversity of the product under test, and that some effects – biological or technological - on the parameter(s) in question have not been investigated in enough detail or are not covered by the databank. This is particularly relevant if one uses data of so-called experimental samples that are authentic but whose production methods might differ from commercial samples. In the example of the EU-Wine databank, the suitability of the data generated using micro-vinification was validated for the interpretation of stable isotope ratios from commercial samples (Christoph, Hermann, & Wachter, 2015).

Non-targeted analysis databases typically relate to spectroscopic/spectrometric “fingerprinting” analyses in which spectra or spectral information are collected. In contrast to the targeted analysis databases, no specific analytical parameters are included, although robust analysis is also paramount. In these cases it is typical for researchers to perform their own statistical analysis and report findings that can be generally applied e.g. the presence of kynurenic acid as a biomarker of sweet chestnut honey (J. A. Donarski, Jones, Harrison, Driffield, & Charlton, 2010). The disadvantage of this approach is that the data cannot easily be recycled for different authenticity problems.

This opinion piece provides what the authors believe are the most important considerations which must be addressed, when creating either an authentic targeted or non-targeted analysis database. It covers three broad sections, relating to aspects that need to be addressed before, during and after the analytical data has been collected. An overview of the recommended steps involved in the collection of a food authenticity database is presented in Fig. 1, the individual steps being covered specifically in more detail in the remainder of the manuscript.

It should be noted that this opinion piece covers only the collection of databases that store analytical data, which can be used to combat a specific subset of fraud types. It is self-understanding that such an approach does not guarantee protection against all types of food fraud that can occur (e.g. theft, quota exceedance, and overrun).

2. Food authenticity database creation

2.1. Definition of the scope of the food authenticity database

The development and maintenance of a food authenticity database can be very resource intensive. Therefore, the most important consideration that needs to be addressed before any sample collection or analysis can begin, is the definition of the specific purpose of a particular database. A database designed without a specific use in mind is likely to overlook considerations that are crucial for specific problems. Conversely, in databases created using non-targeted analysis, given the expense associated with creating them and the increasing use of ‘big-data’, it is prudent to ensure metadata that is not directly relevant to the purpose of the database is also recorded. A balance needs to be reached, such that the recording of additional metadata is not too onerous. For example, where secondary analytical checks are performed on samples before inclusion into any database (e.g. fatty acid profiling of vegetable oils), for example to verify samples of being typical or authentic by established approaches, it is recommended these metadata are included as part of the sample descriptors, where this is possible without additional cost.

The specific primary purpose of the food authenticity database will determine its applicability, and inform what samples need to be analysed and the method in which they will be analysed. The reference samples, which are analysed to create the database, will define its scope. Consideration must be given to assure that samples used to populate the database are comparable to those that will be ultimately challenged by the database. For example, to create a database that is only designed to differentiate between Scottish and English beef (J. Donarski & Heinrich, 2016), representative samples of only English and Scottish beef would be needed to populate the database. Although this is still a considerable task, it can be achieved relatively easily. If the samples are collected only during the period of 2017 though, the scope of the database would only be for beef slaughtered in the year 2017. Therefore, such a database would only be valid to challenge beef samples that are either Scottish or English that were slaughtered in 2017. If this database was challenged with English beef slaughtered in 2018 one of three outputs would occur: the challenge sample would be correctly classified as English, the challenge sample would be erroneously classified as Scottish, or the database would report that the sample could not be classified. In this case, the scope of a database could be increased through assumptions which are then validated. For example, if the factors that lead to the differentiation of English and Scottish beef remain consistent in 2018, then it is highly likely that the database will remain applicable. This can be validated by challenging the database with a limited number of English and Scottish beef samples, slaughtered in 2018. If these samples are correctly predicted, with an accuracy that is not significantly different to samples from 2017, then the assumptions and the use of this particular database in 2018 can be considered as valid.

If this same database was challenged with a sample from a non-English or non-Scottish origin slaughtered in 2017, one of three outputs would occur, that sample would be erroneously classified as English, the sample would be erroneously classified as Scottish, or the database would report that the sample could not be classified. A food authenticity database has no scope for identifying samples that were not considered during its creation, and in this specific example to increase the scope would be a very significant task. As a minimum, it would need to be demonstrated that the factors that led to differentiation between English and Scottish beef samples enabled discrimination between England, Scotland and all other countries. In general, truly global geographical discrimination of food samples may not be achievable. In cases where discrimination of one sample type from all others is possible, these statements can only be considered valid for the samples that have been used to create or challenge the database. It is therefore common for food authenticity databases to be created and used to confirm whether a food is consistent with expected properties. In these cases, a specific property of a sample is measured and suspect samples are challenged against this database, samples that do not fit the expected profile are rejected.

It is vital, that any database is developed with the end user and the intended purpose in mind. Therefore considering at which point of the supply chain the database will be utilised is important. Development of a database where the retailer is the end user, for the purpose of due diligence testing and to act as a deterrent to fraud in the supply chain, will have different requirements to a database where the end user is an international governing body with the purpose of ‘policing’ the food industry. During database planning, it may be beneficial to perform a food supply chain risk assessment to identify where the risk of fraud is highest for a specific product. For example, where the highest risk of fraud is origin (speciality products such as Extra Virgin Olive Oil, products protected by PDO/PGI or products commanding a market

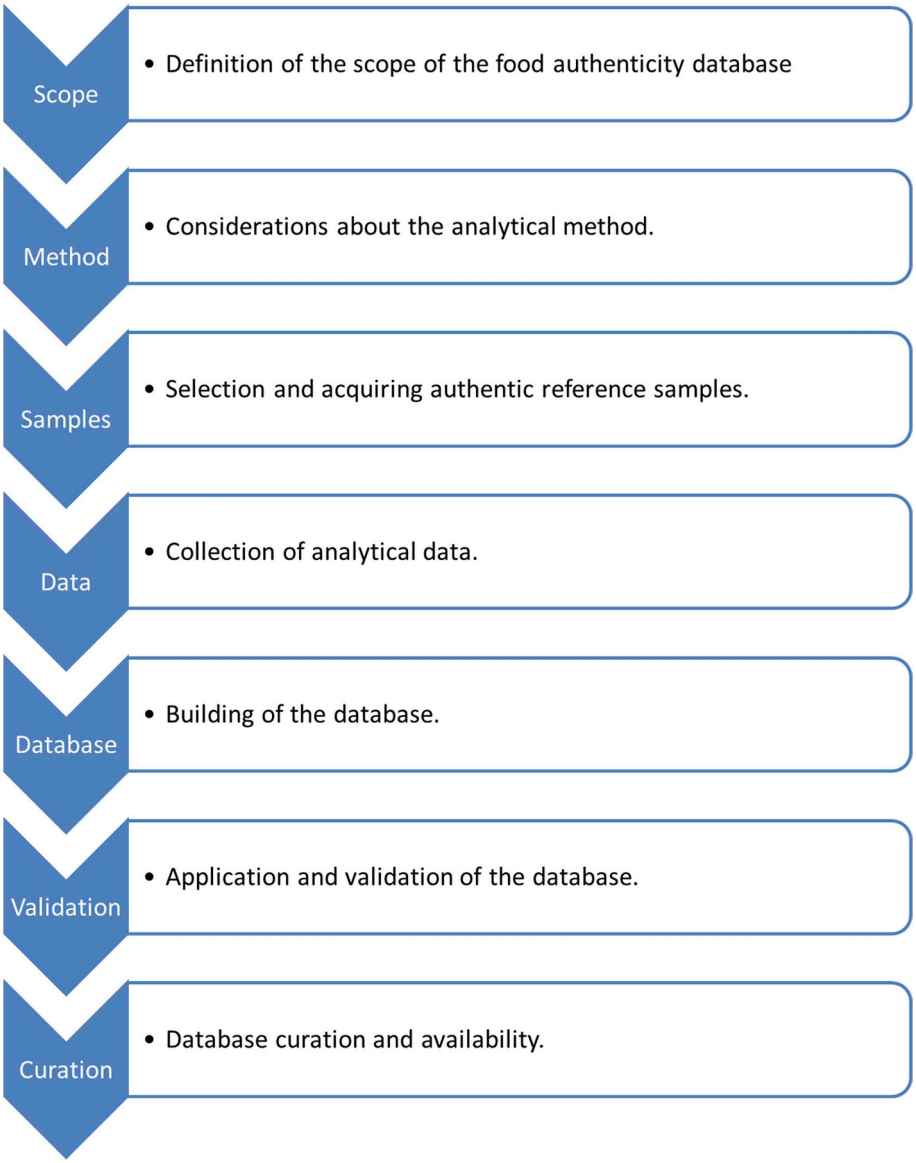

Fig. 1. Flowchart representing the major steps that need to be considered when creating a Food Authenticity Database.

premium such as British Beef) it is necessary to ensure that the database is representative of primary production. Where the risk of food fraud is highest for addition of cheap substitutes or regulated materials during processing it will be necessary for the database to be representative of processed products and may therefore focus more on producers and distributors.

2.2. Considerations about the analytical method

Once the scope of the database has been designed, consideration needs to be given to the analytical method, which will enable discrimination of authentic from non-authentic samples. The most appropriate approach is to consider the scope of the database, and to consult experts in the field of the foodstuff in question to list the physical and chemical distinctions that differentiate the products. Factors to consider include, geographical origin, temperature, age of material, ingredients, production methods and consistency. There is a significant risk that models, created without careful consideration of the analytical differences, will fail. Databases, that seek to differentiate samples from distinct geographical origins, often rely on methods that use some form of stable isotope analysis, as these have shown to be influenced by factors such as altitude and temperature. Databases that seek to differentiate samples by their chemical differences often rely on either targeted or non-targeted methodologies. Non-targeted methods have the advantage of being able at least in principle to also identify new chemical markers, which were originally neither considered nor known.

Considerations need to be given to the analytical method and its long-term ability to produce reproducible data. Stable isotope analysis is a significant field and a range of calibration standards and proficiency test materials are available. Spectroscopic technologies (e.g. NMR, FT-IR, Raman) are considered to be reproducible technologies and are not influenced by drift or changes in sensitivity over time and also have calibration standards available. Hyphenated techniques, such as GC-MS and LC-MS require more consideration. Chromatography columns deteriorate through usage, which can lead to changes in retention times of analytes. MS detectors are also susceptible to fouling, which can impact on signal response. These factors need to be accounted for in analyses and require the use of randomised sample analysis and in-house reference material, which is used for intra and inter batch corrections (Rusilowicz, Dickinson, Charlton, O'Keefe, & Wilson, 2016). More detailed information about methods of analysis is provided in the accompanying scientific opinions produced from the FoodIntegrity

project (Cavanna, Righetti, Elliott, & Suman, 2018; McGrath et al., 2018).

Once a suitable analytical method has been identified, a small scale study to confirm the validity of any assumptions made is recommended. Samples are rarely analysed directly; some sort of extraction is usually required. Therefore, considerations must be given to the reproducibility of the extraction method and the reproducibility of the instrumental method and these should be related to the observed discrimination, which is the scope of the database. If instrument and sample extraction variability can be shown to be minimal, compared to the observed discrimination between groups, replicate extraction of samples and replicate analysis of extracts can be avoided. More detailed information about sampling can be found in the literature (Pawliszyn, 2002).

The act of performing a small study can also highlight any factors that have not been considered (e.g. difficulties in obtaining reference materials).

Identification of the point in the supply chain, where samples should be taken will ensure that the collected samples are fit for the analytical technique being used and the database is representative for the target product. The position in the supply chain, where samples are collected can influence both (i) the quality of the analytical data and (ii) the integrity of the database. For stable isotope ratio analysis, processing or cooking of a raw material and the addition of other ingredients can affect the isotopic composition to an extent that it is no longer comparable to a database of 'raw' or 'unprocessed' material. It is necessary for non-targeted applications to understand the production process of the material of interest. This ensures that the database is representative of expected analytes, which are introduced during production. Dependant on the application, it may be necessary to perform validation to determine the effect of processing on the analyte of interest.

2.3. Selection and acquiring authentic reference samples

Whether building a global database as part of an international collaboration or building a targeted database for internal use in a commercial enterprise, it is necessary to ensure that the database is fit for purpose. Primarily, it means ensuring that the database is representative of the target (authentic) population.

The first, and arguably the most important criteria is that the samples contained within the database are authentic. Inclusion of a fraudulent sample, labelled as an authentic sample, within a food authenticity database will invalidate the database. Extreme care must be taken to procure authentic, relevant reference material. The necessary steps required to acquire authentic reference material differ between commodities, but all stages must be considered when acquiring samples. It should be noted that the resource expended on acquiring authentic reference material is often significant and therefore, once acquired, reference samples are often stored for future alternative uses. In these cases, the samples should be stored in a manner such that the sample is analytically unaltered. This is a non-trivial matter and any stored authentic reference material, should be analytically verified as being unaltered before being reused.

For convenience, purchase of samples from retailers is the easiest way to build rapidly a large dataset. However, the integrity of retail purchased samples is low, as one cannot guarantee authenticity, and therefore the integrity of the created database will be low. Ideally, samples should be collected from primary producers (i.e. farms, fisheries etc.) by impartial collectors (i.e. individuals with no economic incentive to corrupt the database) to ensure that traceability and integrity of reference samples is maintained (Di Egidio, Oliveri, Woodcock, & Downey, 2011). It is important to remember that food fraud is now found at all levels of the food supply chain; if one does not have traceability to the sample's origin, one cannot guarantee its authenticity.

There is a tendency to focus on sample numbers, when considering population quality. The final sample size will be dependent on a number of factors, including (i) access to authentic samples, (ii) project budget, (iii) timeframe, (iv) objectives for the completed database and (v) the logistics of sample collection. A more important consideration during project planning and sample collection is: Does the sample population represent the natural variation of the analyte(s) of interest observed in the target population? Dependent on the question to be answered, it may be necessary to consider natural variation caused by many factors including geographical location, variety or breed, age and health, physical and climate stresses, processing method (e.g. olive oils), temporal or seasonal variation and anthropogenic contamination.

It is also useful to understand the production density of a foodstuff of interest. For example, if considering the origin of tomatoes grown in the UK, it would be non-beneficial to build a database of hundreds of samples grown in Wales, when UK production is predominantly based in the South East. It is necessary to validate the database created, to ensure that it is representative of the target population and fit for purpose by "confirmation by examination and the provision of objective evidence that the particular requirements for a specific intended use are fulfilled" (ISO/IEC 17025:2017. General requirements for the competence of testing and calibration laboratories 2017).

When preparing a sampling strategy, it is also necessary to consider the statistical analysis that is required as this will affect the total number of samples required. Multivariate techniques for non-targeted methods require a sampling size sufficient to build a 'fingerprint' that represents authentic samples (Alewijn, van der Voet, & van Ruth, 2016).

2.4. Collection of analytical data

Once the appropriate analytical method has been chosen and collection of representative samples have been planned or are completed, acquisition of analytical data is required. It has already been discussed that collection of sample metadata should include as a minimum all information relevant to the purposes of the database, and that it is good practice to record other information that is accurately known, under the assumption that this is not a significant administrative burden. Information relating to the specifics of the analytical method should also be reported, such that an expert in the analytical field would be able to exactly recreate the experimental conditions, using comparable equipment. Given the range of analytical methods that can be used to collect analytical data, it is not appropriate to list the minimum reporting information in this document. A range of initiatives have been undertaken to define such a requirement such as those of the Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI)(Sumner et al., 2007).

The physical collection of analytical data should also be considered and follow specific practices. Many experimental factors can influence analytical data that are not directly related to the analysed sample and these should be controlled to ensure that they do not introduce confounding results into the analysis. Examples of confounding effects can include (i) day to day variability of sample extraction, (ii) change in laboratory temperature throughout the working day, (iii) instrument variability and (iv) minor changes in extraction solvent composition. This is most relevant for creation of non-targeted analysis, but it can also impact targeted databases, although such effects are typically covered within the method validation. The most effective method of control is a combination of careful monitoring of known or suspected influences, analysis of a reference sample throughout the database creation and collection of analytical data in a random order, and if needed including the regular collection of such sample (Berg et al., 2013). This will mitigate, monitor and minimise factors that can influence a database. It is typical to either combine a small aliquot of all samples to create an 'in-house reference sample', or to choose a representative sample which is available in sufficient quantity to ensure its extraction and analysis throughout the study, alongside samples of interest.

2.5. Building of the database

When specifying requirements for any database, there are many basic conditions that must be considered. Presently, tools are commercially available which provide useful starting points for database creation so that database builders do not have to consider fundamental requirements. Leveraging such tools allows the researcher to focus on the specific design of the database, e.g. What kind of sample(s)? Which metadata? Which results? Which validation and approval mechanisms?

The fundamental requirements of a database platform can be split, as follows:

2.5.1. Underlying database

The following points for the underlying database need to be considered.

- What storage medium should be used for the underlying database? MySQL, SQL Server, Oracle, MongoDB, etc.
- What has already been used in similar databases and does this matter?
- Is it freely available? Does it need to be?
- How secure does the database need to be?
- Volume of data?
- Backup considerations?
- Speed considerations? Data entry vs retrieval?
- Short, Medium and Long-term goals? Where will the database reside? Will it start local, but end up in the cloud?
- Underlying database schema should be flexible enough to mitigate the need for database builders to modify the schema to accommodate growing metadata or analytical results requirements.

2.5.2. Application layer

The application provides an interface between the database and the user, and streamlines the process of working with the database. The aspects that need to be considered for the application are.

- Security, Logins, Permissions, Auditing (created, updated) The configuration or metadata and results including: Column Name; Data Type; Units; Validation; Categorisation.
- Convenient data entry that provides validation and auto suggestions to prevent “similar” metadata (i.e. an ontology). Depending on the sample record being created, pre-define a list of required fields (i.e. minimum reporting information).
- Approval workflows. E.g. when results have been submitted to the database whether they should be automatically available without review.
- The application layer should, as a bare minimum, allow the data to be retrieved.
- It should be assumed that users and administrators of the database would not be accessing the underlying database directly (for data access or schema changes), but instead access through an application layer.
- A basic requirement is that the data can be read easily through the application layer.
- An application programming interface (API) should be present.
- Support the import of data (e.g. from Excel)
- Support the export of data (migration, backup, etc.)

2.5.3. Database vs application layer (database schema vs application schema)

The database schema should be flexible enough to accommodate many different types of metadata and analytical results (description, units, etc.). To provide enough flexibility, there should be a separation of concerns between the underlying database and its schema and the perceived schema at the application level. Changing the schema at the application level, should not change the schema at the database level. It should be noted that such design can result in a performance reduction, and therefore should only be done if this reduction in performance is not significant and is within acceptable limits of the database application.

Ideally the application layer should go above and beyond these fundamental requirements and in a final database should provide:

- An intuitive framework to aid in the initial design of a database
- Advanced filtering and sorting capabilities (querying)
- Efficient pagination of data (both within the application and API)
- Convenient data visualisations (grids, charts, maps, etc.) and analytics (PCM, LDA, etc.) to quickly sanity check results and perform rudimentary analysis.
- A mechanism to perform grouping and aggregations
- A mechanism to compare and search for similar data sets
- Seamless integration with sample acquisition applications (e.g. IonOS, IonVantage, etc.) via API to streamline data flow.
- Seamless integration with statistics packages to extend database capability (R, SPSS, etc.)

2.6. Application and validation of the database

When applying a database for testing the authenticity of a sample, authenticity limits or profiles will be defined for use in comparison. For targeted analyses, the typical approach is that of univariate data evaluation, based on the calculation of the arithmetic mean, median, standard deviation and the confidence interval considering the Student-Factor (t-distribution) and choosing appropriate confidence levels, e.g. 95% (Camin et al., 2017). Where several other targeted analyses are to be considered, the approach of bivariate evaluation is appropriate. These two approaches can be used only for targeted analysis, whereas for multiparametric targeted or non-targeted databases, a multivariate model is preferable or the only choice.

Suitable multivariate statistics for creating models are described, in detail in the literature (Granato et al., 2018). In brief, multivariate treatment of the data will produce a model able to classify samples as authentic or non-authentic, or into other classes (e.g. different origins). Two stages are normally completed: a data compression step which reduces the size and complexity of the original data, and a modelling step that is carried out on these selected features. The principal statistical methods for creating models are discriminant analysis and class modelling. Discriminant analysis is appropriate for determination of a value (e.g. the oxidation level in olive oil (Lagouri & Gimisis, 2016)), whereas class modelling is used either to define normality for a single class (e.g. is a sample authentic) or for multiple classes (e.g. is an olive oil Italian, Spanish or Greek).

Validation of a database includes both the data within it, and its ability to satisfactorily complete the role for which it was designed. All data used to create the database, must be validated, i.e. reliable. This implies that the laboratory producing the data must demonstrate competence and accuracy. For example, in the case of the EU-Wine databank, laboratories providing data must be accredited according to ISO17025 and must participate in a proficiency test. The best way therefore to validate the data is for laboratories participate in inter-laboratory comparison exercises. Where appropriate, available proficiency tests can be used and when these are not available *ad hoc* round robin tests can be organised. When laboratories upload their data to databases, which are created by several organisations, they should provide indications about their performance in the inter-laboratory comparison exercise in terms of a z-score or deviation from the target value.

For non-targeted databases, the organisation and use of inter-laboratory comparison exercise is very challenging (Riedl, Esslinger, & Fauhl-Hassek, 2015). Hence, metadata describing the measurement device parameters and protocols would be more useful to be included in the database, to demonstrate the reliability of the laboratory and

provided data.

Once the data within the database has been validated, the database must be tested for its validity in being representative. One limited, but often described strategy is to apply resampling of computational methods, the most popular of which is cross-validation. The model or database is built with a test set and the remaining sample set is used to validate it, i.e. to check if model or database is able to recognise the correct grouping of the samples. The validity of the database can also be tested by the use of anonymised samples in duplicate.

More effective and recommended is external validation, i.e. using the database/model to predict a complete independent new set of samples, which should be representative of the population of unknown samples (different years, different production technology, and different analytical instruments), thus mimicking the future use of the database.

During validation the presence of outliers within the data may become apparent. An outlier is an observation, that is well outside of the expected range of values, and can be either discarded from the data set or retained. It could happen that outliers - which are classified as non-authentic - are present in the database and the validation dataset. Outliers can have many reasons: analytical and random errors or they could be the result of a flaw in the assumed theory, calling for further investigation by the researcher. When outliers are present, the first action is to confirm that they are not due to analytical problems or random errors. This is typically determined through re-analysis of the samples in question, and if the re-analysed samples are shown to differ from the original data, these initial results from the samples can be removed as outliers from the database. Where data are removed from the database due to extraction/data acquisition errors, consideration should be given to all other samples analysed within that batch and complete re-analysis of a batch may be required. If on re-analysis of samples, data that are consistent with the initial 'outlier' data are recorded, a further investigation needs to be undertaken to determine the underlying cause. Typically, outliers are due to particular and unique technological or geographical issues, such as a particular microclimate or technological choice. In this case, further investigations are needed to understand if outliers belong to another population of data or if they are just 'outliers' falling in the percentage of error of the chosen confidence level (for example 5% for 95% confidence level). Where outliers are removed from a population within a database for specified reasons e.g. due to geographical issues, the scope of the database might be reduced, and this must be noted so that the database is not erroneously applied.

2.7. Database curation and availability

Following completion of the food authenticity database the curation of the database needs to be considered. The curation of a database is to ensure that the results remain valid over time. This is particularly important where the underlying cause of discrimination of authentic produce/products is not fully understood, and therefore may be affected by seasonality or changes in manufacturing process. It is recommended that authentic and known fraudulent samples that have been produced since the database was created are tested and that the classification results are consistent with those generated during initial validation of the database. Samples that have been guaranteed authentic for database checking can be incorporated into the database, although it is not recommended that challenge samples, even when shown to be authentic through database testing are used to augment the database. The analysis of known, authentic reference material, newly acquired alongside challenge samples will ensure that the database remains valid over time.

When changes in the analytical collection of data occur, for example through replacement of analytical instrumentation or the updating of control software, these need to be considered on a case by case basis. At a minimum it is recommended that a series of known samples, previously analysed and contained within the database, are used to assess the impact of the changes.

Where a database is to be used by multiple parties, consideration needs to be given to database sharing. The comparability of data will have been addressed, but considerations also need to be given to whom access of the database will be granted. It has been shown that those who commit food fraud will adapt their adulterations to produce/products in order to avoid detection.

Therefore, releasing a database for interrogation to all parties may enable fraudsters to adopt their approach. Conversely, databases should be shared such that food is protected. It is recommended that despite the increased administrative burden, a gatekeeper system, where the credentials of those requesting access to a database are evaluated, should be adopted ("Council Regulation (EC) No 510/2006 of 20 March 2006 on the protection of geographical indications and designations of origin for agricultural products and foodstuffs,").

3. Conclusion

The globalisation of foodstuffs brings new and novel commodities to consumers throughout the world. When foodstuffs are new, it can be the case that consumers or even inspection laboratories cannot easily recognise when a fraud has taken place. This, coupled with long supply chains and often minimal deterrents for those that wish to commit fraud, can lead to an increased prevalence of fraud. In these types of fraud those best placed to detect them are often the authorities/regulators in the production area using locally produced food authenticity databases.

The significant investment required to generate a food authenticity database and the sensitivity of the data that they contain, can reduce the willingness of organisations to share their data. Where proprietary food authenticity databases are offered as a commercial service to the food industry, it is recommended that the non-sensitive aspects discussed in this opinion piece are made available to provide confidence that the database is appropriate for its specified use. At the time of publish no single source exists for food authenticity databases although repositories are available (Kale et al., 2016). It is typical for project to create their own individual websites (Álvarez, Pascual, Rusu, & Bogason, 2013) and links to collections of these can be found (e.g. https://www.analyticalresultsdb.com/what-is-ardb/find-databases). It is recommended that a central dedicated repository is created for the storage of analytical data relating to food authenticity. A repository has been developed in the FoodIntegrity project, and its collection of methods and databases (both open and not) will be transferred to the European Union and made open access, it is envisioned that in the future this will act as a central source for data relating to food authenticity.

It is hoped that the recommendations presented in this scientific opinion can assist in the development of databases and the sharing of the data between suitable parties.

Acknowledgements

Funding: This manuscript was produced within the FoodIntegrity Project. The project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 613688.

References

Alewijn, M., van der Voet, H., & van Ruth, S. (2016). Validation of multivariate classification methods using analytical fingerprints – concept and case study on organic feed for laying hens. *Journal of Food Composition and Analysis, 51*, 15–23.

Álvarez, B., Pascual, M., Rusu, A., & Bogason, S. (2013). *A Review on Existing Databases Relevant for Food Fraud and Authenticity*.

Berg, M., Vanaerschot, M., Jankevics, A., Cuypers, B., Breitling, R., & Dujardin, J.-C. (2013). LC-MS metabolomics from study design to data-analysis - using a versatile pathogen as a test case. *Computational and Structural Biotechnology Journal, 4*

e201301002-e201301002.
Camin, F., Boner, M., Bontempo, L., Fauhl-Hassek, C., Kelly, S. D., Riedl, J., et al. (2017). Stable isotope techniques for verifying the declared geographical origin of food in legal cases. *Trends in Food Science & Technology, 61*, 176–187.
Cavanna, D., Righetti, L., Elliott, C., & Suman, M. (2018). The scientific challenges in moving from targeted to non-targeted mass spectrometric methods for food fraud analysis: A proposed validation workflow to bring about a harmonized approach. *Trends in Food Science & Technology, 80*, 223–241.
Christoph, N., Hermann, A., & Wachter, H. (2015). 25 Years authentication of wine with stable isotope analysis in the European Union – review and outlook. *BIO Web of Conferences, 5*, 02020.
Council Regulation (EC) No 510/2006 of 20 March 2006 on the protection of geographical indications and designations of origin for agricultural products and foodstuffs. In.
Di Egidio, V., Oliveri, P., Woodcock, T., & Downey, G. (2011). Confirmation of brand identity in foods by near infrared transflectance spectroscopy using classification and class-modelling chemometric techniques — the example of a Belgian beer. *Food Research International, 44*, 544–549.
Donarski, J., & Heinrich, K. (2016). *BRITISH BEEF ORIGIN PROJECT II – improvement of the British beef isotope landscape map (isoscape) for Scotland and northern Ireland, Vol. 2018*.
Donarski, J. A., Jones, S. A., Harrison, M., Driffield, M., & Charlton, A. J. (2010). Identification of botanical biomarkers found in Corsican honey. *Food Chemistry, 118*, 987–994.
Granato, D., Putnik, P., Kovačević, D. B., Santos, J. S., Calado, V., Rocha, R. S., et al. (2018). Trends in chemometrics: Food authentication, microbiology, and effects of processing. *Comprehensive Reviews in Food Science and Food Safety, 17*, 663–677.
Kale, N. S., Haug, K., Conesa, P., Jayseelan, K., Moreno, P., Rocca-Serra, P., et al. (2016). MetaboLights: An open-access database repository for metabolomics data. *Curr Protoc Bioinformatics, 53*, 11–18 14.13.
Lagouri, V., & Gimisis, A. (2016). 219 - optical non destructive UV-VIS-NIR spectroscopic tools and chemometrics in the monitoring of olive oil phenolic compounds and oxidation. *Free Radical Biology and Medicine, 100*, S102–S103.
McGrath, T. F., Haughey, S. A., Patterson, J., Fauhl-Hassek, C., Donarski, J., Alewijn, M., et al. (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 in Food Science & Technology, 76*, 38–55.
Pawliszyn, J. (2002). Index. In D. Barceló (Vol. Ed.), *Sampling and sample preparation for field and laboratory: Vol. 37*, (pp. 1107–1131). Elsevier.
Riedl, J., Esslinger, S., & Fauhl-Hassek, C. (2015). Review of validation and reporting of non-targeted fingerprinting approaches for food authentication. *Analytica Chimica Acta, 885*, 17–32.
Rusilowicz, M., Dickinson, M., Charlton, A., O'Keefe, S., & Wilson, J. (2016). A batch correction method for liquid chromatography–mass spectrometry data that does not depend on quality control samples. *Metabolomics, 12*, 56.
Smeyers-Verbeke, J., Jäger, H., Lanteri, S., Brereton, P., Jamin, E., Fauhl-Hassek, C., et al. (2009). Characterization and determination of the geographical origin of wines. Part II: Descriptive and inductive univariate statistics. *European Food Research and Technology, 230*, 15.
Sumner, L. W., Amberg, A., Barrett, D., Beale, M. H., Beger, R., Daykin, C. A., et al. (2007). Proposed minimum reporting standards for chemical analysis chemical analysis working group (CAWG) metabolomics standards initiative (MSI). *Metabolomics: Official journal of the Metabolomic Society, 3*, 211–221.