SENSORFINT COST ACTION CA19145

European Network for assuring food integrity using non-destructive spectral sensors (SensorFINT).

Improved algorithms, software, prediction equations and recommendations for data management and NDSS issues

This first document (deliverable 6, released at end of July 2022) sets the topics to be explored and the methodology to assemble the guidelines about “Improved algorithms, software, prediction equations and recommendations for data management and NDSS issues” which will be the final WG3 deliverable/objective.

In practice, this will consist of several steps: the first one is to organize a one-day workshop (end of January 2023) focused on the five WG3 sub-tasks having for each sub-topics a keynote talk in review type, followed by a discussion; second, taking insights and outcomes from the workshop the coordination group will assess state-of-art and methods review, together with main issues/questions; finally, a couple of guideline papers will be delivered (March 2023). Furthermore, outcomes from these activities will integrate into the white paper proposal involving all actions.

⇒ *Topic 1: Sampling strategies aiming at establishing sampling guidelines references*

In the framework of the SensorFint project, the main topics to be investigated include the strategy to define the optimal/minimal number of samples with respect to the specific data analysis task (e.g. geographic origin, adulteration, multivariate calibration, etc.) facing in a correct way the inherent heterogeneity (Topic 1.1); and the need of standard/reference samples for authenticity (Topic 1.2), as an example the possible use of existing food commodities databases and the strategies for data retrieval.

Topic 1.1. The main reference literature concerns declination of general Theory of Sampling (TOS) [1] to food and feed materials [2], recognizing that *a priori* stringent distributional assumptions cannot be adopted and sample heterogeneity (e.g. spatial, temporal, etc.) must be assessed first. In this respect, most of the TOS recommendations and practices hold as well for food and feed. According to these, variographic analysis is the right tool to establish fit-for-purpose sampling protocol for specific commodities and delivery context, as shown in application. Notwithstanding, official guidelines and terminology is still far from adopting a coherent vision in agreement to these principles. Another very relevant aspect in food and feed sampling is the definition of the Sampling Quality Criteria (SQC) [ref. 2 pp. 265-268] which consists of defining the sampling objective, the decision sampling unit (DU) and the required confidence level. In particular, the DU varies with context, such as should it be a single unit (e.g. wheat kernel), a bag, a sac or a track.

In the NDSS context there are additional issues of concern, that would be addressed by WG3:

- Which is the most sensible way to define the DU and cope with SQC when sensors are used in the field (portable sensors and/or drones), e.g. depending on type of crops, fruit cultivation, etc. How the sensor sampling area impacts on the representativeness of sampling, how it could be investigated.

- Main pros. and cons. of spectral and imaging devices from the point of view of sampling representativeness, as well as of design of portable sensors.

- Variogram analysis could need adaptation to the specific cases, especially how to extract features from variograms with irregular trends

- How to treat and/or benefit from sampling replicates (also including the different sampling steps) in model (especially calibration) building. This scenario needs to be exploited considering that use of NDSS allows many replicates while reference measurements do not. A path could be to Investigate resampling methods to enlarge the number of reference measurements

- Seasonality variability should be considered another aspect of sampling, more generally the possible interplay of model updating strategies and varying samples heterogeneity in different sampling campaigns needs to be studied.

Topic 1.2 Modelling authenticity, which in its broad sense includes recognizing compliance to labels and thus could embrace frauds, adulteration, contamination, is particularly challenging from the sampling point of view. In fact, in addition to being representative calibration samples should be truly authentic. This issue poses the theme of if standard reference samples could be available and how they should be assembled. A possible solution is creating food commodities databases under specific guidelines [3]. This is not an easy task, since proper samples collection, storage, verification of samples stability and stability of variability sources with respect to the time authenticity models were built must be taken into deep consideration. These issues are even more critical when models are built based on NDSS and untargeted analysis, research lines which can be addressed:

- developing multivariate strategies to assess stability, e.g. domain applicability concepts, robust methods, etc.

In addition, authenticity database containing NDSS data acquired on representative authentic samples for a given food commodities could serve as basis for further models to be developed, e.g. taking part of the database and integrating with new data, in this case the questions that would be explored are:

- best database mining methodologies
- compatibility of spectral data (sensors data in general) to be integrated with existing data in the database, this issue link to methodology in Topic 4.

⇒ *Topic 2: Fusion of spectra, images and other data (of diverse sensors in general)*

In the framework of the SensorFint project, data fusion can be approached at two levels. The first one (Topic 2.1) consists in merging several spectral sensors, such as NIRS with fluorescence. The second one (Topic 2.2) consists in merging one or more spectral sensors with metadata, e.g. NMR spectra with process data such as temperature, pressure or origin data, vintage data, etc.

Whatever the sub-topic, data fusion can be done following different strategies, as detailed in Azcarate et al. [4]. In general, three types of fusion levels are reported (low-, mid-, and high-level) [5]. Low-level fusion involves using block-level information directly in the development of the model. It can be done by simply concatenating the blocks and using monoblock methods. It can also use specific methods of decomposition or factorization of a block with respect to others [6]. Mid-level fusion begins with a step of extracting features from each dataset, using statistical analyses such as PCA and PLS. The scores are then merged by simple concatenation and fed into a classical single-block method [6]. In high-level fusion, single-block models are developed on each block. Then, the outputs of these models are merged into a final model [6]. This high-level fusion is also called stacking. When low level fusion is achieved, we are in the case of multi block analysis.

- A general question would be addressed by WG3, about the theoretical properties of each existing multi block method.
- Especially, discussing about the methods

Topic 2.1 is closely related to WG2, whose objective is recalled below:

WG2 is aiming at exploring the potential of combining several NDSS for solving critical food integrity issues which cannot be solved using one sensor alone. This integration will enable collection of information about composition and distribution, microbial contamination, etc., using hyperspectral imaging combined with reflectance or fluorescence as examples. This task will investigate the fusion or integration of signals from several sensors (NIR, hyperspectral imaging, fluorescence, Raman, and others), to provide new advantages and challenges in addressing food quality, authenticity, and safety problems, unsolved by sensors of any single type.

A special focus will be on increasing the potential for inspection of exogenous contaminants, as well as detecting intrinsic changes in food products resulting from contamination and/or changes in thermal conditions during processing such as overheating, incomplete drying, etc. It is important to highlight the scientific breakthroughs and practical potential related to the development of technologies that allow integration of different non-destructive spectral sensors and their implementation in different parts of the food chain.

Topic 2.1 of WG3 will therefore be devoted to the study of chemometrics tools to implement the above WG2 objectives. Some literature exists on the application of spectral data fusion to food characterization and integrity, e.g. [7-9]. A deeper review should however be done to understand the potential and the limits of NDSS data fusion for food. Some practical and scientific issues are already stated:

- Level of data fusion: Besides simply testing the different fusion levels (low, medium and high), the question arises from a signal processing point of view. What are the theoretical arguments for each of the fusion levels, when the data to be fused are spectra?

- Block definition: While it seems natural to put each spectrum type in a different block, is this the best strategy?
- Fusion method: Regardless of the level of fusion, which methods are most suitable for spectra?
- Spectral images: How to process spectral images in a data fusion process?

Topic 2.2 aims to explore a less trivial way of data fusion than Topic 2.1, and moreover not mentioned in the SensorFint project. It is a question of making the most of all available data and thus, to study the fusion of spectral data with other data, of different types, coming from the process, in the broad sense. It can be measurements made on-line of an industrial process, such as a pressure, a flowrate, or a meta data related to the origin of the sample, such as the variety, the season. The literature is much less abundant for this topic than for the previous one, at least for food applications. In [10], a recent comprehensive review shows that most of the information fused to the spectra comes from non-destructive sensors, such as noses, tongues or e-noses. In [11, spectra are fused with vision data. If we broaden the application domain, we can find interesting references in polymer chemistry [12, 13] or petrol chemistry [14].

A deeper review should thus be done to understand the potential and the limits of fusion between NDSS and meta data / process data for food. Some practical and scientific issues are already stated:

- Level of data fusion: Besides simply testing the different fusion levels (low, medium and high), the question arises from a chemometrics point of view: is the type of fusion dependent from the type of meta data?
- Is it necessary to select the variables in the meta data / process variable bloc?
- In the case of low-level fusion, what kind of multi block method is the most suited?
- As for Topic 2.1, what is the influence of the block definition? Is it interesting to split / merge blocks?
- Investigation of the potential of machine learning / deep learning

⇒ *Topic 3: In-process NDSS real-time analysis (food industry, big/SME)*

NDDS are particularly suitable for on/in-line implementation in the food processing industry as well as for continuous in-situ/field monitoring, this brings to the issue of developing models for real-time analysis, monitoring and control. The real-time scenario [15] poses several issues from the modelling point of view: models, in the monitoring phase, should be computationally fast, i.e. share the same rate as the data acquisition rate; should adapt to process drifts (process "normal" dynamic); should distinguish between "pure" sensor data variability (or drift, e.g. caused by spectral source aging, dirtiness, etc.) not connected to changes in process and/or intermediate product quality. This latter aspect is linked, in the case of spectral/imaging sensors, also to the preprocessing issue [14]. Study integrating process variables with spectroscopic sensors are still limited [12-13] and most often at laboratory or pilot scale. Moreover, a large part of the literature studies deals with end product quality property predictions and rarely with developing multivariate monitoring charts. The salient research questions that would be addressed:

- Model adaptation in real time, investigate and compare local modelling approaches [16] with recursive or adaptive ones [17], like those based on updating the covariance matrix [18], as well as on the fly methodologies [19].
- Review fault detection strategies in MSPC and assess eventual specificity arising for spectral sensors (e.g. how to solve the sensors drift issue, etc.)
- Complementarity/integration with deep learning approaches (DL)
- Interplay with data fusion and data feature selection (see Topic 2)
- Consider different scenarios, such as:
  - o levels at which computation/model implementation take place at the sensor; locally at the factory; or on the cloud (this question is connected to WG4 activities)
  - o time scale of the monitoring, e.g. industrial context (e.g. sensors on conveyor belts, fermentation tank, etc.) or in-field.

⇒ *Topic 4: Cloning instrument, model maintenance and calibration transfer*

Most literature on the application of NDDS reports proofs of concept that are limited to the calculation of a model (calibration) and its application on a so-called independent data set (validation, or test). However, the use of NDDS also requires proving that the performances obtained during this first validation remain valid when conditions change. This generic problem is referred to in chemometrics as robustness. When the measurement conditions of a spectrum change, the measured spectrum x is added with a deviation, dx. The reproducibility of the model, and thus of the sensor, with respect to this deviation, defines the robustness.

SensorFint project covers a large range of processes, and thus deviation sources. Instrument cloning, whether between laboratory instruments or to a dedicated process instrument is certainly the most concerning issue for the deployment of an NDDS-based application [20-21]. The most common approach to solving this type of problem is to measure standard samples on both devices and then apply optical normalization methods, such as PDS [22], or orthogonalization, such as TOP [23]. Changes in acquisition conditions, such as temperature, particle size, moisture, constitute another category of problems, for which other methods have been developed. They are based on the observation of dx as a function of the variations of the influence factors, then on the implementation of a correction strategy using the measured dx [24]. Changes in origin, species, and growing or harvesting conditions are another category of causes of non-robustness. They are characterized by the fact that it is impossible to measure standard samples, or even to measure dx. In [25], the DOP method was proposed to create virtual standards from a few reference values. This method is also particularly suitable to compensate for drifts in online measurements of unknown origin [26]. Finally, the trickiest case arises when the conditions change, but the only available data are spectra acquired in both conditions. This problem, better known in machine learning as domain adaptation, has been recently studied in chemometrics [27-28]. Although all these issues have been addressed separately over the last twenty years, they all fall under the same topic, which can be called model maintenance. The related research questions are:

- the inventory and comparison of existing methods, with a view to the applications of the SensorFint project

- the development of a unified view of model maintenance, incorporating the machine learning perspective
- the implementation of these methods in the framework of an NDDS network

⇒ *Topic 5: Validation (model level, long term, etc..) aiming at establishing validation guidelines*

Validation is the founding stone of every modelling task and a core concept in chemometrics, so it is already implicitly included in all topics above. In food quality and authentication modelling by NDSS a special focus is on the fact that all steps from raw data acquisition to the long-term use of proposed models have to be validated. In general, two main validation steps can be highlighted:

- from sampling to data acquisition (sampling design, sample preparation, measurement, one or more platforms, data acquisition, and assembly)

This is included in Topic1.1, and mostly important sampling issues are not to be seen independently from the second validation step below. In fact, validation at modelling phase requires that the raw data acquired for new/future samples, i.e. validation samples, will be affected by the same sources of variability with respect to the calibration data, and the analytical measurements are of the same quality. The latter aspects may not be trivial with NDSS.

- from raw data to model use (data preprocessing/pretreatment, model building, model diagnostic and validation, model interpretation, eventual model refinement, and model stability). These issues are strictly linked to Topic 4, but also implicit when evaluating Topic 2-3.

References:

1. Gy, P.M. (1998) Sampling for Analytical Purposes, John Wiley & Sons Ltd, Chichester, UK

2. Esbensen et al., Representative Sampling for Food and Feed Materials: A Critical Need for Food/Feed Safety, Journal of AOAC International , 2015, 98 (2), pp. 249 - 320.

3. J. Donarski, F. Camin, C. Fauhl-Hassek, R. Posey, M. Sudnik, Sampling guidelines for building and curating food authenticity databases, Trends in Food Science & Technology 90 (2019) 187–193.

4. Azcarate, S. M., Ríos-Reina, R., Amigo, J. M., & Goicoechea, H. C. (2021). Data handling in data fusion: methodologies and applications. TrAC Trends in Analytical Chemistry, 143, 116355.

5. Hall, D. L., & Llinas, J. (1997). An introduction to multisensor data fusion. Proceedings of the IEEE, 85(1), 6-23.

6. Cocchi, M. (Ed.). (2019). Data fusion methodology and applications. Elsevier.

7. Biancolillo, A., Boqué, R., Cocchi, M., & Marini, F. (2019). Data fusion strategies in food analysis. In Data handling in science and technology (Vol. 31, pp. 271-310). Elsevier.

8. Borràs, E., Ferré, J., Boqué, R., Mestres, M., Aceña, L., & Busto, O. (2015). Data fusion methodologies for food and beverage authentication and quality assessment–A review. Analytica Chimica Acta, 891, 1-14.

9. Márquez, C., López, M. I., Ruisánchez, I., & Callao, M. P. (2016). FT-Raman and NIR spectroscopy data fusion strategy for multivariate qualitative analysis of food fraud. Talanta, 161, 80-86.

10. Zhou, L., Zhang, C., Qiu, Z., & He, Y. (2020). Information fusion of emerging non-destructive analytical techniques for food quality authentication: A survey. TrAC Trends in Analytical Chemistry, 127, 115901.

11. Di Rosa, A. R., Leone, F., Cheli, F., & Chiofalo, V. (2017). Fusion of electronic nose, electronic tongue and computer vision for animal source food authentication and quality assessment–A review. Journal of Food Engineering, 210, 62-75.

12. De Oliveira RR, Avila C, Bourne R, Muller F, de. Juan A. Data fusion strategies to combine sensor and multivariate model outputs for multivariate statistical process control. Anal Bioanal Chem 2020;412(9):2151–63. https://doi.org/10.1007/s00216-020-02404-2.

13. Strani L, Mantovani E, Bonacini F, Marini F, Cocchi M. Fusing NIR and Process Sensors Data for Polymer Production Monitoring. Front Chem 2021;9:748723. https://doi.org/10.3389/fchem.2021.748723.

14. Buendia Garcia, Lacoue-Negre, M., J., Gornay, J., Mas Garcia, S., Bendoula, R., & Roger, J. M. (xxx). A novel methodology for determining effectiveness of preprocessing methods in reducing undesired spectral variability in near infrared spectra. under review

15. P. Kadlec, B. Gabrys, S. Strandt, Data-Driven Soft Sensors in the process industry. Computers and Chemical Engineering 33 (2009) 795–814.

16. Matthieu Lesnoff, Maxime Metz, and Jean-Michel Roger. Comparison of locally weighted PLS strategies for regression and discrimination on agronomic NIR data. J. Chemometrics, 34(5) 2020. doi:10.1002/cem.3209.

17. P. Kadlec et al. Review of adaptation mechanisms for data-driven soft sensors, Computers and Chemical Engineering 35 (2011) 1–24.

18. Dayal, B. S., & MacGregor, J. F. Recursive exponentially weighted PLS and its applications to adaptive control and prediction. Journal of Process Control, 1997, 7(3),169–179.

19. Vitale, R., Zhyrova, A., Fortuna, J., de Noord, O., Ferrer, A. & Martens, H. On-The-Fly Processing of continuous high-dimensional data streams. Chemometr. Intell. Lab. 2017, 161, 118-129 .

20. Shenk, J. S., Westerhaus, M. O., & Templeton Jr, W. C. (1985). Calibration transfer between near infrared reflectance spectrophotometers 1. Crop science, 25(1), 159-161.

21. Folch-Fortuny, A., Vitale, R., De Noord, O. E., & Ferrer, A. (2017). Calibration transfer between NIR spectrometers: New proposals and a comparative study. Journal of Chemometrics, 31(3), e2874.

22. Wang, Y., Veltkamp, D. J., & Kowalski, B. R. (1991). Multivariate instrument standardization. Analytical chemistry, 63(23), 2750-2756.

23. Andrew, A., & Fearn, T. (2004). Transfer by orthogonal projection: making near-infrared calibrations robust to between-instrument variation. Chemometrics and Intelligent Laboratory Systems, 72(1), 51-56.

24. Chauchard, F., Roger, J. M., & Bellon-Maurel, V. (2004). Correction of the temperature effect on near infrared calibration—application to soluble solid content prediction. Journal of near infrared spectroscopy, 12(3), 199-205.

25. Zeaiter, M., Roger, J. M., & Bellon-Maurel, V. (2006). Dynamic orthogonal projection. A new method to maintain the on-line robustness of multivariate calibrations. Application to NIR-based monitoring of wine fermentations. Chemometrics and Intelligent Laboratory Systems, 80(2), 227-235.

26. 26: Zeaiter, M., Latrille, É., Gras, P., Steyer, J. P., Bellon-Maurel, V., & Roger, J. M. (2022). Improvements in the Robustness of Mid-Infrared Spectroscopy Models against Chemical Interferences: Application to Monitoring of Anaerobic Digestion Processes. AppliedChem, 2(2), 117-127.

27. Nikzad-Langerodi, R., Zellinger, W., Lughofer, E., & Saminger-Platz, S. (2018). Domain-invariant partial-least-squares regression. Analytical chemistry, 90(11), 6693-6701.

28. Fonseca Diaz, V., Rogerb,J.M., Saeys, W., (2022). Unsupervised Dynamic Orthogonal Projection. An efficient approach to calibration transfer without standard samples. (under review).

NOTE: In addition to the working document produced, it agreed to prepare some publications related to the topics highlighted in a Special a Special Issue in the TrAC (Trend in Analytical Chemistry) Journal The papers submitted, except one of them, are still under revision. https://www.sciencedirect.com/special-issue/10L0Q2G73ZS