



Rapid selection of milk mid-infrared spectra for creating a dairy cow population world representative spectral database

H. Soyeurt,^{1*} S. Franceschini,¹ M. Bahadi,² J. Leblois,³ Y. Brostaux,¹ F. Dehareng,⁴ M. Frizzarin,⁵ K. Tiplady,⁶ L. Dale,⁷ and C. Nickmilder¹

¹TERRA Teaching and Research Centre, Gembloux Agro-Bio Tech, University of Liège, 5030 Gembloux, Belgium

²Lactanet, Sainte Anne de Bellevue, QC H9X 3R4, Canada

³Walloon Breeders Association, Elevéo, 5590 Ciney, Belgium

⁴Agricultural Walloon Research Centre, 5030 Gembloux, Belgium

⁵Teagasc, Animal & Grassland Research and Innovation Centre, Moorepark, Fermoy P61 P302, Co. Cork, Ireland

⁶Livestock Improvement Corporation, Hamilton 3240, New Zealand

⁷Regional Association for Performance Testing in Livestock Breeding of Baden-Wuerttemberg (LKV Baden-Wuerttemberg), 70190 Stuttgart, Germany

ABSTRACT

The advantage of employing mid-infrared spectrometry for milk analysis in breeding lies in its ability to quickly generate millions of records. However, these records may be biased if the calibration process does not account for their spectral variability when constructing the predictive model. Therefore, this study introduces a novel method for developing a world representative spectral database (WRSD) to reduce the risks of spectral extrapolation when predicting dairy traits in new samples. Using a 2-phase selection procedure that is both efficient and minimizes memory usage, we first generate a decomposition matrix via principal component analysis (PCA) on a dataset of 2,324,443 records. The next phase iterates spectral selection based on a location index from PCA scores, calculating spectra occurrence frequency for refined barycenter estimations. The chosen spectra's barycenter closely aligns with the entire dataset, proving the efficacy of using just 3 principal components. Applied to 4 varied datasets totaling over 21 million records, we select 583,440 spectra to represent spectral diversity, with selection rates between 2.00% and 14.88%. This selection illustrates the spectral variability across different dairy populations and data providers. Demonstrated through a hypothetical calibration set of 71 samples, the WRSD's utility for algorithm developers becomes apparent. This calibration set covers between 91.42% and 98.50% of the WRSD variability, except for the Irish dataset (3.50%), indicating a need for additional samples to accurately represent Irish variability and minimize spectral extrapolation. This study offers valuable

insights into the representativeness of training sets for capturing spectral variability within targeted dairy populations. Although the current WRSD does not fully encompass global milk spectral diversity, its development underscores the importance of gathering more data and standardizing spectral information across spectrometer brands. Ultimately, the WRSD proves crucial not just for trait prediction but also for identifying abnormal milk samples, also marking a significant relevance in dairy science technology.

Key words: milk, mid-infrared, equation

INTRODUCTION

Mid-infrared (MIR) spectroscopy has been a longstanding method for characterizing the composition of milk. Recent research indicates that in addition to merely quantifying fat and protein content, this technique can be employed to directly or indirectly estimate various compounds associated with milk composition. These estimations prove valuable in evaluating the nutritional quality, technological attributes, animal welfare considerations, and the environmental impact of milk (Soyeurt, 2023). The insights derived from these predictions find application in the development of decision-support tools and the formulation of animal selection programs.

The predictive models are formulated by algorithms, commonly referred to as equations, which are applied to MIR spectra obtained from milk samples collected during routine procedures by DHI organizations or as part of milk payment processes between the dairy producers and the dairy companies. Numerous equations are routinely used in practice, often predicting the same trait. However, a significant challenge arises in the absence of a unified validation set, preventing a direct comparison of these equations. A validation set is crucial for

Received March 13, 2024.

Accepted June 27, 2024.

*Corresponding author: hsoyeurt@uliege.be

The list of standard abbreviations for JDS is available at adsa.org/jds-abbreviations-24. Nonstandard abbreviations are available in the Notes.

assessing the reliability of predictions yielded by an equation.

Reliability fluctuations can occur for various reasons, including inadequacies in modeling, the inability of MIR to accurately predict the target trait, disparities in the measurement of reference data between calibration and validation sets, or oversight in accounting for spectral variability within a population (Grelet et al., 2021). Notably, the absence of spectral variability in the calibration set leads to spectral extrapolation during the prediction phase, introducing an additional bias. To address this challenge, it proves relevant to calculate the standardized Mahalanobis distance, also known as the global-H (GH) distance, between the sample to be predicted and the calibration set's barycenter (Zhang et al., 2021).

Nevertheless, the automated estimation of the GH distance remains unimplemented in several dairy laboratories, despite the undeniable value of this approach (Zhang et al., 2021). The reluctance to embrace systematic GH distance calculation for each prediction might stem from some equation suppliers hesitating to disclose information regarding the characteristics of their calibration sets. Thus, a potential workaround to this challenge lies in estimating the spectral coverage encompassed by an equation's calibration set within the animal population targeted for the application of the equation.

To achieve this, a comparison between the calibration set derived from the equation supplier and a spectral dataset representative of the target population is essential. Because only spectral information is required, and not the reference values used to construct the equation, one could anticipate a higher willingness on the part of equation suppliers to collaborate. Additionally, establishing such a spectral database representative of the spectral variability within a population would prove beneficial for developing unsupervised approaches aimed at identifying anomalous spectra in milks or animals, indicative of abnormal milk or ailing cows. For instance, Franceschini et al. (2022) demonstrated the potential to detect cows with potential metabolic disorders using such an approach. The creation of a representative spectral database not only facilitates equation refinement, but also (thanks to the limited size of the dataset) opens avenues for innovative unsupervised methodologies geared toward identifying abnormal milks or animals based on spectral irregularities, offering broader applications beyond equation validation.

Given the outlined interests, proposing a method for constructing a representative spectral database becomes imperative. Various approaches can be considered, each presenting distinct advantages and drawbacks. Initially, we examine the divide-and-conquer approach, as outlined by Franceschini et al. (2022). This methodology involves

partitioning the dataset into multiple subsets, followed by the application of a clustering method to determine distinct groups within these subsets. Subsequently, a representative average spectrum, referred to as a centroid, is generated for each group. These centroids are then amalgamated to form a final clustering representing the entire population. The use of subsets is crucial to optimize the computational cost of this calculation. Constructing a clustering necessitates computing the distance between all samples within the subset, a process that can swiftly consume a substantial amount of memory. Despite this, the advantage of the approach lies in its capacity to indirectly account for all samples in a dataset. However, a notable disadvantage is that the final clustering no longer relies on real spectra, potentially introducing bias. Furthermore, in our specific case, having multiple representatives for the same spectral variability, as permitted by this approach, is undesirable. Consequently, this approach proves unsuitable for our study.

An alternative approach involves selecting samples based on their GH distance, a method employed by Soyeurt et al. (2011) to identify the most pertinent samples for a prediction equation. The underlying concept is that not all samples with a low GH distance should be considered, as this indicates their proximity. Although this selection aligns more closely with our requirements, constructing the distance matrix poses a challenge due to the extensive datasets available. To circumvent the creation of an overly large matrix, a strategy involves calculating the GH distance only in relation to specific iteratively selected samples. In the initial round, a sample is randomly chosen from the dataset, and GH distances are computed exclusively in reference to this sample. Subsequently, samples close to it are eliminated from the dataset. In subsequent rounds, a new sample is randomly selected from the remaining pool, and GH distances are recalculated relative to this new sample. This process is repeated until no close samples, i.e., those with a low GH, remain. Although this method appears efficient in theory, its computational demands are substantial and may result in prolonged calculation times. Additionally, to mitigate collinearity among spectral points for distance calculation, a principal component analysis (PCA) must be executed on the dataset. This, too, presents a computational challenge in terms of memory, particularly when dealing with datasets that encompass millions of records.

Upon reflection, conducting a PCA on the entire dataset may not be imperative. Alternatively, envisioning the application of PCA to a smaller dataset and subsequently extending the loadings to the entire database emerges as a viable strategy. Although this approach may not yield a flawless decomposition into orthogonal axes, it becomes

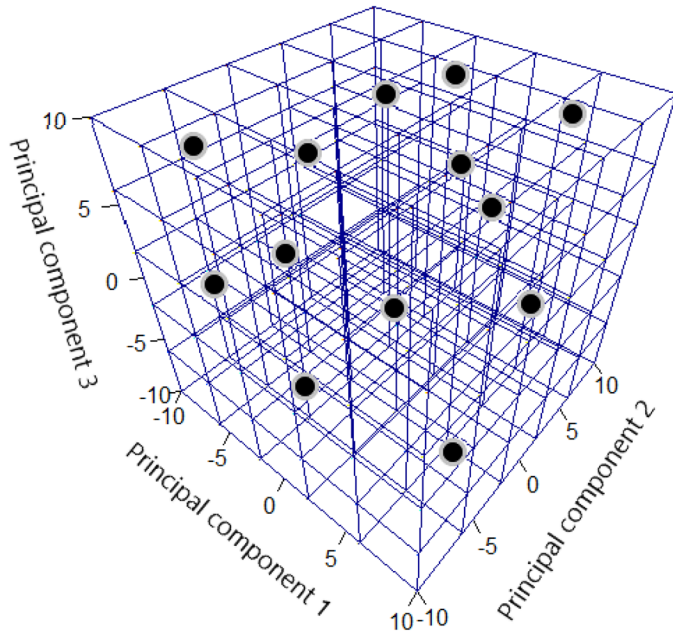


Figure 1. Representation of sample proximity using a principal component analysis.

inconsequential if our primary concern is projecting all samples onto these axes. The coordinates on the principal components for each sample inherently serve as a direct indicator of sample proximity, as depicted in Figure 1. In other words, similarly to a photography discretizing a scenery continuous signal into an ensemble of pixels, this methodology allows to discretize the signal variability into a set of elements equivalent to hypervoxels in a n -dimensional space (the PCA space) whose axes represent different components of MIR signal variability. In the current study, our objective is to validate the suitability of this method for selecting samples that adequately represent the existing spectral variability within a dairy cow population.

MATERIALS AND METHODS

This research is structured into 2 distinct parts. The initial phase revolves around the formulation of the decomposition matrix derived from a specific set of samples, followed by an investigation into the representativeness of the spectra chosen through the identified principal components. Subsequently, the second segment of the research applies the initial findings to various milk recordings, encompassing raw milk spectra collected from Foss (Hillerød, Denmark) MIR spectrometers. This application aims to evaluate the variability within these records and highlight the significance of the proposed approach in mitigating spectral extrapolation during the prediction process.

Decomposition Matrix

To construct the PCA decomposition matrix, we used a subset of the Walloon spectral database created by Elevéo (Ciney, Belgium) that pertains to milk recording in southern Belgium. This dataset encompasses a substantial 2,324,443 records, with the dataset size strategically chosen to optimize our server's capacity for performing PCA. Concretely, the selection was done to have an entire year. We chose the year 2021, and then records were chosen randomly to have more than 2,000,000 records. The final dataset comprised records from 351,060 cows across 2,010 farms, spanning from January 2017 to November 2021.

In the preprocessing stage, a first derivative was applied to the raw spectral data to rectify the baseline of the spectral signal. This derivative, configured with a window of 5 spectral points, was followed by the removal of noisy regions in the spectrum, in accordance with the methodology outlined by Vanlierde et al. (2021). The used spectral areas were situated between wavenumbers 968 and 1,577, 1,720 and 1,809, and 2,561, and 1,966 cm^{-1} , resulting in a focus on 289 spectral points. The PCA was then performed on the normalized spectral points using the `prcomp` function from the “stats” package (version 3.6.2) in R version 4.3.2 (R Foundation for Statistical Computing). The eigenvectors, means, and standard deviations, employed for normalizing the data, were saved in external files for subsequent application to new data.

Selection of Representative Spectra

In the initial phase, we applied the selection of representative spectra to the Elevéo subset to determine the optimal number of principal components required for the selection process. In the subsequent phase, this selection method is applied to spectra collected from other milk recording organizations or research institutes.

This selection process is split into 2 distinct parts to facilitate parallelization, with all codes developed in Python 3.8.

In the first part, a script is employed to compute the coordinates of the principal components (PC) for each sample, as defined in the preceding decomposition phase. Subsequently, to identify samples of interest, these coordinates are rounded up to the nearest integer. By concatenating these rounded scores, a sample location index is generated within the space defined by the PC scores. For instance, a sample with a first PC score of 12 and a second PC coordinate of -2 is assigned a location index such as “12-2.” To expedite the computation of coordinates on the PC, the program subdivides the initial file into 10 subfiles. The calculation is then

Table 1. Characteristics of the datasets used to select the representative spectra

| DHI organization or research institute | Country | Brand | Instrument | Total records | Selected spectra | % Selection |
|--|---------|-------|------------|---------------|------------------|-------------|
| Elevéo | Belgium | Foss | MilkoScan | 7,702,328 | 177,045 | 2.30 |
| Lactanet | Canada | Foss | MilkoScan | 10,033,266 | 201,115 | 2.00 |
| Conseil Elevage 25-90 | France | Foss | MilkoScan | 3,096,927 | 86,460 | 2.79 |
| Teagasc | Ireland | Foss | MilkoScan | 798,385 | 118,820 | 14.88 |
| Total | | | | 21,630,879 | 583,440 | |

executed in parallel by processing 5 subfiles simultaneously, leveraging Python's multiprocessing library.

In the second part, the selection of spectral samples unfolds as follows. The process begins by placing the first sample in the dataset, along with its corresponding location index, at the initial position of the vector related to the selected samples. Subsequently, in an iterative manner, the program examines the location indices of the already selected samples. If the index of the current sample under scrutiny is absent from the list of selected samples, it is added. Conversely, if the location index is already present in the vector, the sample is disregarded. Consequently, samples sharing the same index are considered akin. To mirror the distribution of samples within the population, the frequency of occurrence for each index in the database is computed. This density information is then used to weigh the selected spectra, thereby deriving a barycenter that accurately represents the population. This barycenter was calculated as follows:

$$barycenter_i = \frac{\sum_{j=1}^{j=number_selected_spectra} spectrum_{j,i} \times density_j}{\sum_{j=1}^{j=number_selected_spectra} density_j},$$

where i is the spectral point, j corresponds to the number of selected spectra, $spectrum$ represents the value of the first derivative obtained for sample j at spectral point i , and $density$ corresponds to the number of spectra having the same location index. The barycenter from the full database was estimated by calculating the average for each weighted spectral point.

In contrast to an approach reliant on a distance matrix, the method employed here involves only one calculation to estimate the location indices. As more PC scores are incorporated into the location index, minor variations in the spectrum are taken into account, consequently influencing the number of selected samples. To gauge the number of PC scores needed for a representative barycenter, this study will experiment with location indices constructed using a single PC score or up to a certain number of PC scores. The endpoint of this range will be the last PC explaining up to 1% of the spectral variability; this threshold is commonly considered as a tipping point for noise filtering. For each of these combinations,

the study will analyze the number of selected spectra and the variability observed in the obtained barycenters. A paired Student's t -test will then be conducted to assess the similarity between the real and reconstituted barycenters. This comprehensive evaluation aims to determine the optimal number of PC scores necessary to achieve a representative barycenter.

Practical Application

As mentioned previously, the objective of this second phase of the current study is to apply the decomposition of Elevéo subset and the sample selection methodology to real spectral databases obtained from DHI organizations and research centers, using a total of 4 distinct datasets, the characteristics of which are outlined in Table 1.

We aimed to illustrate the significance of a world representative spectral database (**WRSD**) for algorithm developers. To achieve this, a fictitious dataset of 71 spectra (also derived on 5 points window-width and with 289 spectral points selected) was used as a training set for constructing a prediction equation. The subsequent goal was to estimate the proportion of spectral variability covered by this training set within the WRSD or its constituent sets. To achieve this, we calculated the GH distances from WRSD samples to the barycenter of the training set. The GH value between a sample spectrum in the WRSD or its constituents and the calibration barycenter was determined using the formula outlined by Zhang et al. (2021):

$$GH_i = \left[(\bar{x}_i - \bar{\mu})^T S^{-1} (\bar{x}_i - \bar{\mu}) \right] / nPC,$$

where i is one spectral sample in the WRSD; \bar{x}_i is the PC score of the i spectrum to be predicted, $\bar{\mu}$ is the mean of PC scores estimated from the calibration set, S corresponds to the (co)variance matrix between PC scores estimated from the calibration spectra, and nPC is the number of PC used. Next, we tallied the number of samples with a GH distance exceeding 3, indicative of being too distant from the training set and thus likely to yield predictions resulting from extrapolation. Subsequently, the count of outliers was done to estimate the percentage

Table 2. Number of selected spectra using 1 to 4 principal components, as well as the part of the spectral variability explained by those principal components, the *P*-values of the paired Student's *t*-test conducted from the weighted barycenters, and the statistics of the barycenter differences

| Item | Samples | | Paired <i>t</i> -test <i>P</i> -value | PCA ¹ | | Difference | | |
|------|---------|-------|--|------------------|--------|------------|----------|----------|
| | N | % | | % Cumul. exp. | % Exp. | Minimum | Mean | Maximum |
| PC1 | 231 | 0.01 | 0.17 | 48.12 | 48.12 | 1.11E-06 | 3.34E-04 | 3.26E-03 |
| PC2 | 8,741 | 0.38 | 0.26 | 65.25 | 17.13 | 1.92E-07 | 3.21E-04 | 3.88E-03 |
| PC3 | 111,936 | 4.82 | 0.35 | 77.02 | 11.77 | 2.27E-07 | 2.45E-04 | 2.34E-03 |
| PC4 | 534,404 | 22.99 | 0.81 | 81.71 | 4.68 | 4.00E-10 | 1.53E-04 | 1.19E-03 |

¹% Cumul. exp. = percentage of cumulative explained variance; % exp = percentage of explained variance.

of samples with a GH value below 3. This percentage serves as a representation of the dataset's coverage of spectral variability and was calculated as follows:

$$Spectral_{coverage} = 100 - \frac{100}{N_{sample}} \times N_{outliers},$$

where *Nsample* is the total number of samples in the WRSD or its constituent sets; *Noutliers* is the number of spectral samples with a GH distance higher than 3. This assessment of spectral variability coverage provides crucial insights for a developer, indicating whether their training set accurately captures the spectral variability present in the target dairy population where they aim to implement their algorithm.

RESULTS AND DISCUSSION

Selection of Representative Spectra

The PCA was conducted on a dataset comprising slightly over 2 million records (i.e., subset of Elevéo, Table 1). Preceding the analysis, the data underwent derivation, extraction of noise variables, and standardization. The results revealed that 10 and 21 components explained 95% and 99% of the spectral variability, respectively. Beyond the ninth PC, each subsequent component contributed less than 1% to the explained variability. Therefore, various sample location indices, constructed from 1 to 9 PC scores, were employed to select spectra from the studied Elevéo subset. The number of selected samples rapidly increases. Hence, Table 2 displays results solely for location indices constructed from 1 to 4 PC, as the intention is to avoid selecting the entire content of the database. Using a location index composed of the rounded scores for the first 4 PC results in the selection of approximately a quarter of the spectra.

However, upon visual inspection, the pattern of the barycenter constructed from the selected spectra does not exhibit noticeable differences from that derived from the entire dataset, as illustrated in Figure 2. Even with the score of a single PC, there are no significant distinc-

tions between the barycenter derived from the complete database and the one derived from the selection of 231 samples (Table 2). Although not statistically significant, the results in Table 2 reveal that an increase in the number of PC in the location index diminishes the observed differences between the barycenter calculated from the full dataset and the weighted one constructed from the selected spectra, as expected. It is worth noting that despite the absence of significant differences, there is a degree of heterogeneity within the barycenter differences. This variation is illustrated in Figure 3, showcasing the absolute differences observed between the full barycenter and those reconstructed. Although the absolute differences are generally low, an observable heterogeneity is present, particularly in the representation of the last spectral points. This discrepancy is anticipated, given that the first component, explaining the majority of spectral variability, predominantly captures major elements in milk predicted by the initial spectral points, as demonstrated by Soyeurt et al. (2010). By employing a location index based on 3 or 4 PC, a more accurate reflection of the terminal spectral region is achieved, proving particularly useful for quantifying fatty acids (Grelet et al., 2015).

In pursuit of constructing a comprehensive global database representative of dairy spectral variability, the selection of an exceedingly large number of data items is not practical. For instance, the number of selected spectra increases more than 4-fold between 3 and 4 PC. Therefore, considering that the location index derived from the scores of 3 PC adequately captures the terminal zone of the spectrum (differences lower than 0.25%), we suggest establishing this limit.

Practical Application

This methodology for selecting representative spectra can be applied to any spectral database. In the current study, we employed 4 different datasets. The number of selected spectra per dataset using a location index based on 3 PC scores is provided in Table 1. Notably, we found no apparent correlation between the total number of records and the number of selected spectra. The percentage

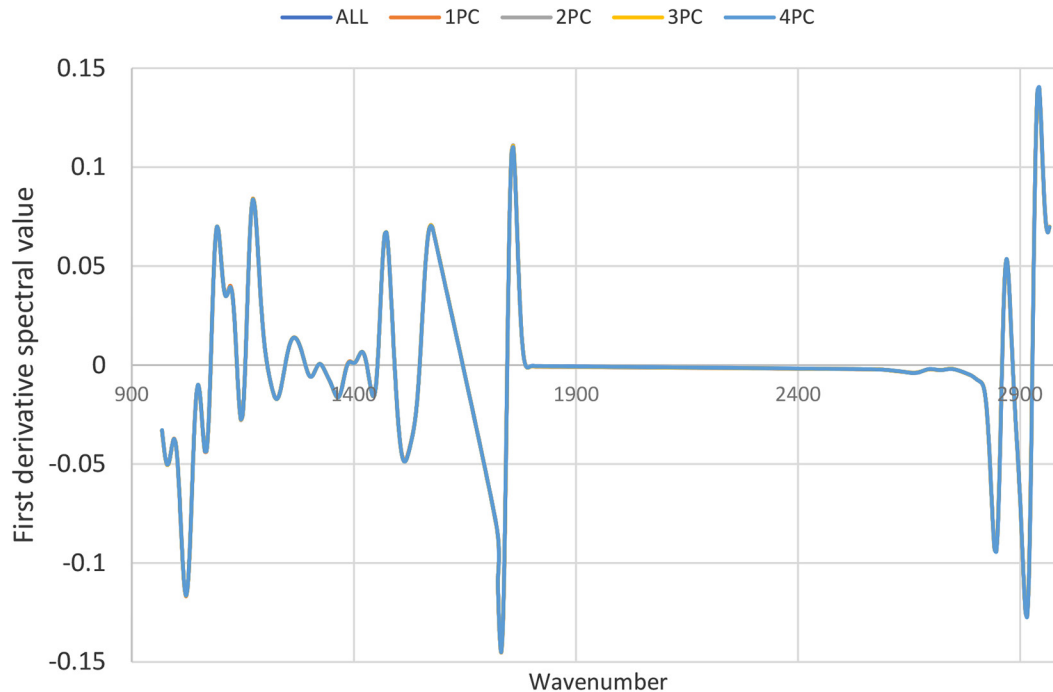


Figure 2. Barycenter obtained with the complete database (ALL) or after selection of spectra using the localization index combining 1, 2, or 3 principal components (PC).

of selected spectra varies from 2.30% to 14.88%. This outcome is anticipated, as the selection is more influenced by the spectral variability present in the dataset. Consequently, the Teagasc (Fermoy, Ireland) dataset exhibits greater spectral variability. The Teagasc dataset included data from 2008 to 2020 from different research farms. Across the different farms, various experiments were carried out, such as the impact of grazing different swards on milk production (McClearn et al., 2019) or of experiencing different stocking rates (Coffey et al., 2018). The majority of the cows were grass-based; however, the indoor system was executed in one research herd. Differences in cow genetic merit existed across the different farms, and different breeds (i.e., Holstein-Friesian, Jersey, as well as cross-breed with Norwegian red) were raised. Moreover, differences across years can be also expected within the same research herd, as the grass-based system is strongly affected by weather conditions. Therefore, different weather conditions can bring changes in milk production (Zhang et al., 2020) and in the spectra.

By consolidating all records into a unified database, the WRSD is formulated. The points denote selected spectra originating from Foss instruments ($n = 583,440$ records). As expected, our method also encompasses the selection of extreme spectral records. The GH distances of samples estimated from the barycenter of this WRSD, using the 8 PC explaining 95% of the variability, ranged

from 0 to 46.81. Samples with a GH higher than 5 were omitted, representing very extreme cases. Consequently, the cleaned WRSD consisted of 580,500 records, constituting a deletion of 0.50%. This cleaned WRSD is depicted in Figure 4.

A comprehensive understanding of the existing variability in milk samples is crucial for identifying abnormal milk samples (Ceniti et al., 2023). Calculating the distance between the barycenter of the WRSD or one of its constituents enables the estimation of whether a sample falls within the population range. Moreover, for predictions aimed at minimizing spectral extrapolation, it is imperative to assess whether the training set, also known as the calibration set, adequately covers the variability specific to the region or country where the prediction algorithm will be applied. Figure 5 provides a visual representation of potential insights. To illustrate this concept, consider a scenario where a developer possesses a training set consisting of 71 samples. To ensure the barycenter calculation is representative for each country, it is based on the density of each selected spectrum.

From our fictitious training set, it is observed that 243,780 WRSD spectra had a GH distance greater than 5, representing 41.99% of the dataset. However, when considering the density of each selected spectrum, this accounts for only 5.35% of the spectral variability. In other words, the samples selected from the calibration set covered almost 95% of the variability present in the

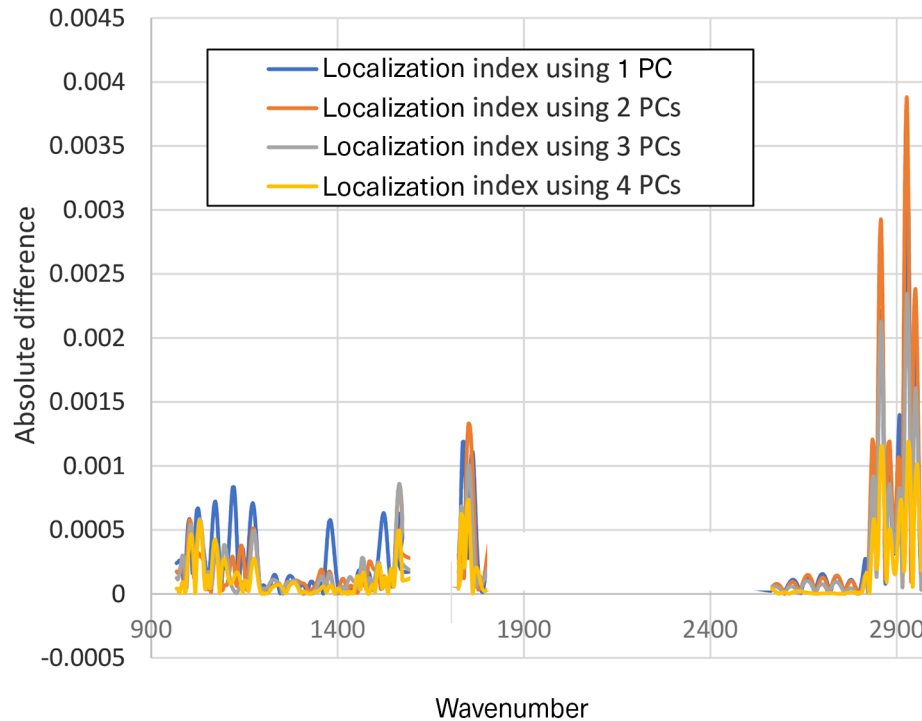


Figure 3. Absolute differences in first derivative spectral values between the barycenter estimated over the entire population or via selected spectra.

spectral databases. We can further examine the constituents of this database, which are the data providers (refer to Table 3). As previously observed with the entire WRSD, examination per data provider underscores the importance of factoring in the frequency of occurrence of the selected spectra.

For instance, using the Lactanet (Sainte Anne de Bellevue, Canada) dataset as an example, without weighting by the frequency of occurrence, the conclusion would be that the calibration set covers only 39% of the variability. However, when weighting the spectra by their frequency of appearance in the Lactanet database, it is evident that the chosen calibration set covers 91% of the variability, particularly in the most represented spectral zone of this population. Conversely, for the Teagasc dataset (which includes mainly data from pasture-based animals), there is little difference, possibly indicating that the spectra in the calibration set are too distant from the typical spectral variability recorded by this research center. This underscores the need to increase the number of samples in the database to encompass the spectral variability existing in this geographical area; otherwise, there is a risk of conveying biased predictions due to potential spectral extrapolation.

The methodology proposed in this study is also applicable to spectra collected from Bentley (Chaska, MN)

instruments. To demonstrate this, spectral data were collected from Germany ($n = 10,776,814$) and New Zealand ($n = 2,044,094$). In accordance with the proposed methodology, a PCA was initially performed on the New Zealand dataset. The PCA loadings for the first 3 principal components were then applied to both the German and New Zealand datasets. Spectral selection was based on the scores obtained for these 3 PC, similar to the process used for the Foss datasets in this study. From the New Zealand dataset, 81,078 spectra were selected, and 89,805 spectra were selected from the German dataset. These selected spectra were then combined to form the WRSD illustrated in Figure 6. Due to differences in spectral wavenumbers, it was not possible to use the PCA loadings obtained from the Elevéo subset. Therefore, a separate WRSD had to be created. Nevertheless, this poses a challenge due to variations in spectral resolution among brands. Consequently, conducting the same analysis for each brand, including the creation of a PCA based on spectral data from the specific brand, becomes necessary.

The known spectral instability and resolution differences within instruments and across brands can be addressed by implementing routine standardization of spectral data, correcting discrepancies within and between instruments. Looking ahead, the adoption of an

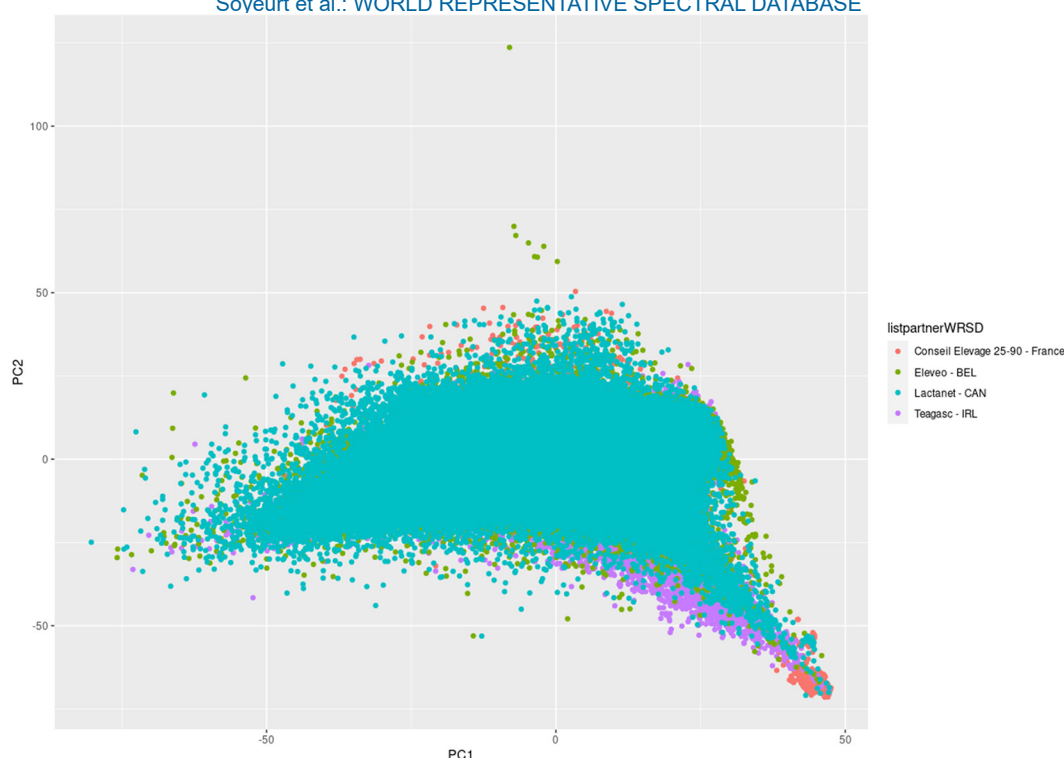


Figure 4. Cleaned world representative database constructed from 4 datasets including milk Foss spectra and illustrated using the first 2 principal components (PC). BEL = Belgium; CAN = Canada; IRL = Ireland.

international standardization approach, akin to the one proposed by Grelet et al. (2017), could eliminate the need for separate decomposition matrices. Although alternative methodologies have been proposed, such as Bonfatti et al. (2017), they come with the inconvenience of standardizing only spectrometers belonging to the same brands.

CONCLUSIONS

The current study introduces a methodology for creating a WRSD, aiming to mitigate spectral extrapolation issues that may arise when predicting a trait on new milk samples using an equation. This representative spectral

database can also be valuable in identifying abnormal milk samples. The proposed method is efficient, and the memory allocations required for the process are limited, particularly due to the parallelization potential offered by the 2-step selection procedure. However, it is essential to acknowledge that the WRSD developed in this study does not fully represent the world's milk spectral variability. Achieving this goal requires data from numerous DHI and dairy organizations, a pursuit aligned with the objectives of the ExtraMIR project managed by the International Committee for Animal Recording and the International Dairy Federation. To create a single WRSD, it is imperative to ensure comparability of spectra across machines and brands.

Table 3. Percentage of spectral coverage of the developer calibration set on the constituents of the world representative spectral database¹

| Item | WRSD without density | | | WRSD with density | | |
|--------------------------------|----------------------|-----------|-----------|-------------------|-----------|-----------|
| | Nsamples | Noutliers | % Covered | Nsamples | Noutliers | % Covered |
| Conseil Elevage 25-90 (France) | 85,995 | 23,877 | 72.23 | 3,093,721 | 46,339 | 98.50 |
| Eleveo (Belgium) | 176,184 | 109,906 | 37.62 | 7,695,533 | 564,446 | 92.66 |
| Lactanet (Canada) | 200,283 | 122,435 | 38.89 | 10,028,709 | 860,240 | 91.42 |
| Teagasc (Ireland) | 118,038 | 112,326 | 4.84 | 797,093 | 772,601 | 3.07 |

¹Nsamples = total number of samples in the WRSD or its constituent sets; Noutliers = number of spectral samples with a GH distance higher than 3.

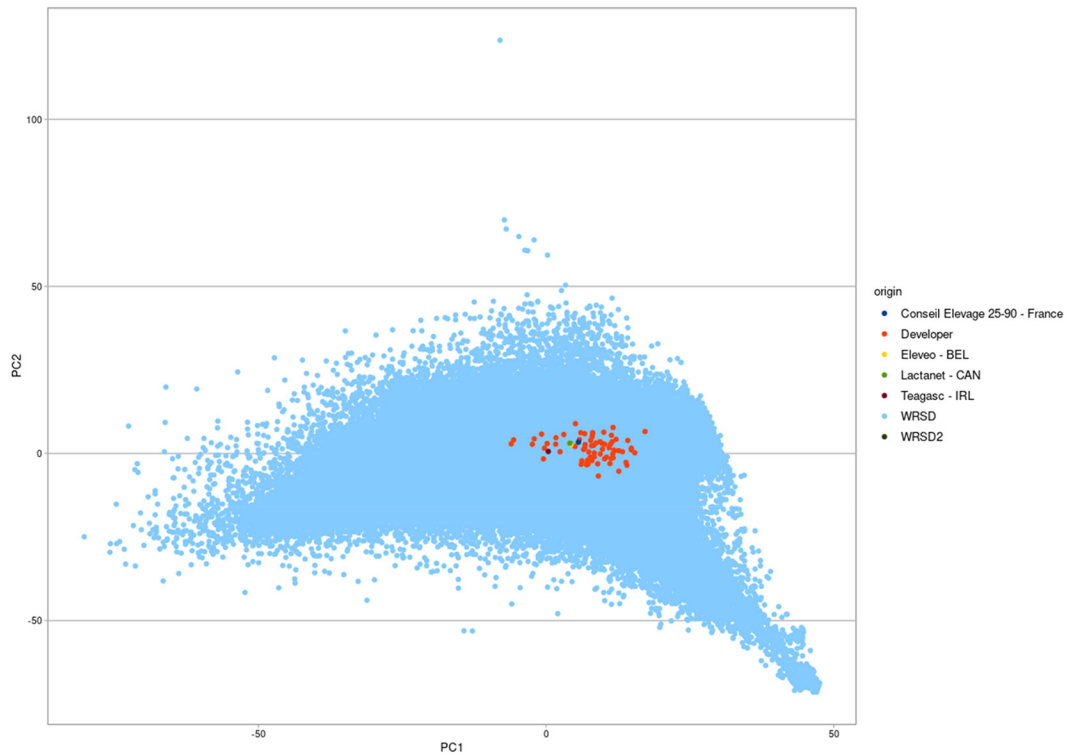


Figure 5. Representation of the training set of the developer in the WRSD spectral space. The weighted barycenters of all WRSD constituents are shown in different colors. BEL = Belgium; CAN = Canada; IRL = Ireland.

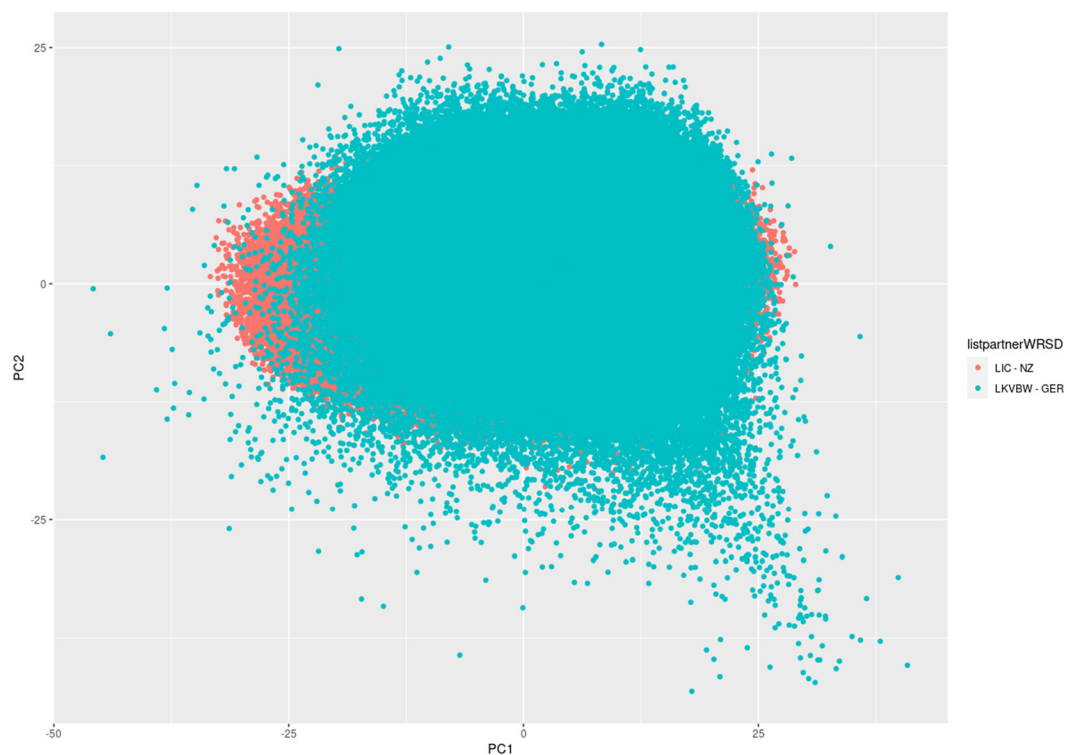


Figure 6. World representative database constructed from 2 datasets including milk Bentley spectra and illustrated using the first 2 principal components (PC). NZ = New Zealand; GER = Germany; LIC = Livestock Improvement Corporation; LKV BW = LKV Baden-Wuerttemberg.

NOTES











We acknowledge INTERREG NWE (Brussels, Belgium; grant agreement no. NWE0100132) as well as the Walloon Region (Belgium) for their financial support in the Holicow project. The authors acknowledge the ExtraMIR project coordinated by the International Committee for Animal Recording (ICAR; Brussels, Belgium) and the International Dairy Federation (IDF; Brussels, Belgium). The existence of this project ensured the networking needed to realize the proposed work. Valérie Wolf (Conseil Elevage 25-90, Roulans, France) is acknowledged for providing representative spectra. No human or animal subjects were used, so this analysis did not require approval by an Institutional Animal Care and Use Committee or Institutional Review Board. The authors have not stated any conflicts of interest.

Nonstandard abbreviations used: % cumul. exp. = percentage of cumulative variance; % exp = percentage of explained variance; BEL = Belgium; CAN = Canada; GER = Germany; GH = global-H; IRL = Ireland; LIC = Livestock Improvement Corporation; LKVBW = LKV Baden-Wuerttemberg; MIR = mid-infrared; Noutliers = number of spectral samples with a GH distance higher than 3; Nsamples = total number of samples in the WRSD or its constituent sets; NZ = New Zealand; PC = principal component; PCA = principal component analysis; WRSD = world representative spectral database.

REFERENCES

- Bonfatti, V., A. Fleming, A. Koeck, and F. Miglior. 2017. Standardization of milk infrared spectra for the retroactive application of calibration models. *J. Dairy Sci.* 100:2032–2041. <https://doi.org/10.3168/jds.2016-11837>.
- Ceniti, C., A. A. Spina, C. Piras, F. Oppedisano, B. Tilocca, P. Roncada, D. Britti, and V. M. Morittu. 2023. Recent advances in the determination of milk adulterants and contaminants by mid-infrared spectroscopy. *Foods* 12:2917. <https://doi.org/10.3390/foods12152917>.
- Coffey, E. L., L. Delaby, C. Fleming, K. M. Pierce, and B. Horan. 2018. Multi-year evaluation of stocking rate and animal genotype on milk production per hectare within intensive pasture-based production systems. *J. Dairy Sci.* 101:2448–2462. <https://doi.org/10.3168/jds.2017-13632>.
- Franceschini, S., C. Grelet, J. Leblois, N. Gengler, and H. Soyeurt. 2022. Can unsupervised learning methods applied to milk recording big data provide new insights into dairy cow health? *J. Dairy Sci.* 105:6760–6772. <https://doi.org/10.3168/jds.2022-21975>.
- Grelet, C., P. Dardenne, H. Soyeurt, J. A. Fernandez, A. Vanlierde, F. Stevens, N. Gengler, and F. Dehareng. 2021. Large-scale phenotyping in dairy sector using milk MIR spectra: Key factors affecting the quality of predictions. *Methods* 186:97–111. <https://doi.org/10.1016/j.ymeth.2020.07.012>.
- Grelet, C., J. A. Fernández Pierna, P. Dardenne, V. Baeten, and F. Dehareng. 2015. Standardization of milk mid-infrared spectra from a European dairy network. *J. Dairy Sci.* 98:2150–2160. <https://doi.org/10.3168/jds.2014-8764>.
- Grelet, C., J. A. F. Pierna, P. Dardenne, H. Soyeurt, A. Vanlierde, F. Colinet, C. Bastin, N. Gengler, V. Baeten, and F. Dehareng. 2017. Standardization of milk mid-infrared spectrometers for the transfer and use of multiple models. *J. Dairy Sci.* 100:7910–7921. <https://doi.org/10.3168/jds.2017-12720>.
- McClearn, B., T. J. Gilliland, L. Delaby, C. Guy, M. Dineen, F. Coughlan, and B. McCarthy. 2019. Milk production per cow and per hectare of spring-calving dairy cows grazing swards differing in *Lolium perenne* L. ploidy and *Trifolium repens* L. composition. *J. Dairy Sci.* 102:8571–8585. <https://doi.org/10.3168/jds.2018-16184>.
- Soyeurt, H. 2023. Fourier transform mid-infrared milk screening to improve milk production and processing. *JDS Commun.* 4:61–64. <https://doi.org/10.3168/jdsc.2022-0294>.
- Soyeurt, H., F. Dehareng, N. Gengler, S. McParland, E. Wall, D. P. Berry, M. Coffey, and P. Dardenne. 2011. Mid-infrared prediction of bovine milk fatty acids across multiple breeds, production systems, and countries. *J. Dairy Sci.* 94:1657–1667. <https://doi.org/10.3168/jds.2010-3408>.
- Soyeurt, H., I. Misztal, and N. Gengler. 2010. Genetic variability of milk components based on mid-infrared spectral data. *J. Dairy Sci.* 93:1722–1728. <https://doi.org/10.3168/jds.2009-2614>.
- Vanlierde, A., F. Dehareng, N. Gengler, E. Froidmont, S. McParland, M. Kreuzer, M. Bell, P. Lund, C. Martin, B. Kuhla, and H. Soyeurt. 2021. Improving robustness and accuracy of predicted daily methane emissions of dairy cows using milk mid-infrared spectra. *J. Sci. Food Agric.* 101:3394–3403. <https://doi.org/10.1002/jsfa.10969>.
- Zhang, F., J. Upton, L. Shalloo, P. Shine, and M. D. Murphy. 2020. Effect of introducing weather parameters on the accuracy of milk production forecast models. *Inf. Process. Agric.* 7:120–138. <https://doi.org/10.1016/j.inpa.2019.04.004>.
- Zhang, L., C. Li, F. Dehareng, C. Grelet, F. Colinet, N. Gengler, Y. Brostaux, and H. Soyeurt. 2021. Appropriate data quality checks improve the reliability of values predicted from milk mid-infrared spectra. *Animals (Basel)* 11:533. <https://doi.org/10.3390/ani11020533>.

ORCIDS

- H. Soyeurt  <https://orcid.org/0000-0001-9883-9047>
- S. Franceschini  <https://orcid.org/0000-0001-6298-5149>
- M. Bahadi  <https://orcid.org/0000-0001-7210-1861>
- J. Leblois  <https://orcid.org/0000-0001-5112-0262>
- Y. Brostaux  <https://orcid.org/0000-0001-6172-7869>
- F. Dehareng  <https://orcid.org/0000-0002-6733-4334>
- M. Frizzarin  <https://orcid.org/0000-0001-7608-5504>
- K. Tiplady  <https://orcid.org/0000-0002-3307-9208>
- L. Dale  <https://orcid.org/0000-0002-5165-4720>
- C. Nickmilder  <https://orcid.org/0000-0001-5235-2145>