INTERNATIONAL STANDARD

ISO 21543

IDF 201

Second edition
2020-10

Milk and milk products — Guidelines for the application of near infrared spectrometry

Lait et produits laitiers — Lignes directrices pour l'application de la spectrométrie dans le proche infrarouge

Reference numbers
ISO 21543:2020(E)
IDF 201:2020(E)

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva
Phone: +41 22 749 01 11

Email: copyright@iso.org
Website: www.iso.org

International Dairy Federation
Silver Building • Bd Auguste Reyers 70/B
B-1030 Brussels
Phone: +32 2 325 67 40
Fax: +32 2 325 67 41
Email: info@fil-idf.org
Website: www.fil-idf.org

Published in Switzerland

Contents

Page

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular, the different approval criteria needed for the different types of ISO documents should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives).

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO shall not be held responsible for identifying any or all such patent rights. Details of any patent rights identified during the development of the document will be in the Introduction and/or on the ISO list of patent declarations received (see www.iso.org/patents).

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation of the voluntary nature of standards, the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT), see www.iso.org/iso/foreword.html.

This document was prepared by Technical Committee ISO/TC 34, *Food products*, Subcommittee SC 5, *Milk and milk products*, and the International Dairy Federation (IDF). It is being published jointly by ISO and IDF.

This second edition cancels and replaces the first edition (ISO 21543 | IDF 201:2006), which has been technically revised. The main changes compared with the previous edition are as follows:

— the measurement principles "Transmittance" and "Transflectance" have been added and defined;

— all sample types have been covered: liquids, solids and semi-solids;

— the calibration and validation sections have been reviewed and updated;

— the outlier section has been revised and the plots renewed;

— the procedures for sample handling and measurement have been expanded to include liquid samples and other examples;

— Annex A has been expanded to include raw milk analysis references.

Any feedback or questions on this document should be directed to the user's national standards body. A complete listing of these bodies can be found at www.iso.org/members.html.

IDF (the International Dairy Federation) is a non-profit private sector organization representing the interests of various stakeholders in dairying at the global level. IDF members are organized in National Committees, which are national associations composed of representatives of dairy-related national interest groups including dairy farmers, dairy processing industry, dairy suppliers, academics and governments/food control authorities.

ISO and IDF collaborate closely on all matters of standardization relating to methods of analysis and sampling for milk and milk products. Since 2001, ISO and IDF jointly publish their International Standards using the logos and reference numbers of both organizations.

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. IDF shall not be held responsible for identifying any or all such patent rights. Details of any patent rights identified during the development of the document will be in the Introduction and/or on the ISO list of patent declarations received (see www.iso.org/patents).

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

This document was prepared by the IDF *Standing Committee on Statistics and Automation* and ISO Technical Committee ISO/TC 34, *Food products*, Subcommittee SC 5, *Milk and milk products*. It is being published jointly by IDF and ISO.

The work was carried out by the IDF/ISO Action Team (S16) of the *Standing Committee on Statistics and Automation* under the aegis of its project leader, Mr A. Niemoeller (DE).

Milk and milk products — Guidelines for the application of near infrared spectrometry

1 Scope

This document gives guidelines for the use of near infrared (NIR) spectrometry in the analysis of milk and milk products in liquid, semi-solid or solid form. Depending on the sample form and application, different instrument setups for transmittance, diffuse reflectance or transflectance can be applied.

2 Normative references

There are no normative references in this document.

3 Terms and definitions

For the purposes of this document, the following terms and definitions apply.

ISO and IEC maintain terminological databases for use in standardization at the following addresses:

— ISO Online browsing platform: available at https://www.iso.org/obp

— IEC Electropedia: available at http://www.electropedia.org/

3.1
near infrared spectroscopy
NIR spectroscopy
spectroscopic method for measuring spectra using the NIR region of the electromagnetic spectrum (from 780 nm to 2 500 nm, 12 800 cm^{-1} to 4 000 cm^{-1})

3.2
near infrared spectrometry
NIR spectrometry
application of *near infrared spectroscopy* (3.1) yielding quantifiable results or other application-related evaluations of NIR spectra

3.3
parameter
constituent
proximate (e.g. total solids, moisture, non-fat solids, fat, protein, lactose and salt contents) or other property (e.g. pH value) of samples applicable for NIR analysis methods

Note 1 to entry: Most parameters are measured in mass fractions expressed in per cent, but some parameters will be determined in percentage by volume or with specific units, e.g. pH and freezing point depression.

4 Principle

4.1 General

A NIR method contains the following steps:

— a sample is presented to a NIR instrument;

— if required, it is preprocessed in advance to obtain a homogeneous test sample representing the chemical composition of the sample material;

— it is loaded into the sample holder of the NIR spectrometer and, during the spectroscopic measurement, the absorbance at wavelengths in the NIR region is measured;

— the spectral data are transformed to constituent concentrations by calibration models developed on samples from the population to be tested (see 7.1).

4.2 Measurement principles

4.2.1 Transmittance (transmission)

In transmittance measurements, light is directed at a sample with a focused or parallel beam. The transmitted part is detected after passing through the sample. This setup is not only used for clear liquids, but also scattering samples in liquid form (e.g. milk, cream) or semi-solid form (e.g. yoghurt, cheese) can be analysed by diffuse transmittance with a suitable optical setup. The optical path length is an important parameter and should be optimized and maintained according to the manufacturer's recommendations.

4.2.2 Diffuse reflectance (reflection)

When light is directed onto a sample, it is reflected and transmitted at differing amounts depending on the bulk properties of the material. Solid and semi-solid samples reflect light in diffuse reflection, i.e. reflecting in all angles equally or near-equally distributed. There are different instrumental setups used to collect and detect diffuse reflected light. Instruments with two principles for illuminating the sample in a container and collecting the reflected light either from above (top-down set-up) or from the bottom through a window (bottom-up) are common.

4.2.3 Transflectance (transflection)

When partially transparent samples are analysed, the reflectance and transmittance properties are not sufficient for using one of these dedicated measuring principles. By selecting a suitable path length and having a reflecting surface (reflector, mirror) on the other side of the sample, the incident light is transmitted through the sample, reflected and transmitted backwards to be detected. Thus, the diffuse reflected and the transmitted light is measured in a ration depending on the optical properties of the sample.

In cases where an NIR instrument provides reflectance measurements, even transparent or turbid liquids can be analysed in transflectance.

5 Reagents

Use only reagents of recognized analytical grade, unless otherwise specified, and distilled or demineralized water or water of equivalent purity.

5.1 Ethanol, or another appropriate solvent or detergent mixture, for cleaning re-usable sample containers.

5.2 Cleaning solutions.

For analysis of liquids in flow cells, appropriate cleaning solutions and maintenance protocols should be used according to the manufacturer's recommendations.

6 Apparatus

6.1 Near infrared (NIR) instrument.

NIR instruments are based on diffuse reflectance, transmittance or transflectance measurement principle by NIR spectroscopy in the whole NIR wavelength region of 780 nm to 2 500 nm (12 800 cm^{-1} to 4 000 cm^{-1}) or segments of this or at selected wavelengths.

The optical operation principle may be dispersive (e.g. grating monochromators), interferometry or non-thermal (e.g. light-emitting diodes, laser diodes, lasers). The instrument should be provided with a diagnostic test system, e.g. for testing photometric instrument noise, wavelength/wavenumber accuracy and wavelength/wavenumber precision/repeatability (for scanning spectrophotometers) according to the manufacturer's recommendations.

The instrument should be equipped with sample containers (cups, glass or polystyrene Petri dishes, flow cells, etc.), which allow repeatable filling of the sample. It is important that the window, Petri dish or flow cell material that allows the NIR light to interact with the sample is manufactured to a consistent standard by the manufacturer or supplier to avoid biased results or even re-calibration work. A sufficient sample volume or surface is required to eliminate any influence from inhomogeneity derived from the chemical composition or physical properties of the test sample. The sample path length (sample thickness) in transmittance measurements should be optimized according to the manufacturer's recommendations with respect to signal intensity for obtaining linearity and maximum signal/noise ratio. In reflectance measurements in a top-down instrument set-up, a quartz window or other appropriate material to eliminate drying effects should preferably cover the interacting sample surface layer. In a bottom-up instrument setup, the container and the sample are covered and protected against drying by the measurement window.

The sample containers may be re-usable or made of disposable material.

6.2 Grinding or grating device.

An appropriate device for preparing the sample (e.g. a food processor for semi-hard cheese) can be required. A procedure should be established to ensure consistency over time and with different operators.

Changes in grinding or grating conditions may influence the NIR measurements and results.

7 Calibration and initial validation

7.1 Selection of calibration samples

The instrument should be calibrated before being used or installed pre-calibrated. Because of the complex nature of NIR spectral data, which consists mainly of overtones and combination bands of fundamental vibrations in the mid-infrared region, the instrument should be calibrated using a series of natural samples (often at least 120 samples). No specific procedure for the calibration of a specific application can be given due to the wide variations over all aspects in sample types, constituents, concentration ranges, instrument setups and possible calibration algorithms. Examples and results are published in an IDF bulletin[1].

The accuracy and robustness of calibration models are dependent on the strategies used for sample selection and calibration. Developed calibration models are only valid for samples covered by the range and variations of the calibration samples. The first step in calibration development is therefore to define the application (e.g. sample types and concentration ranges). When calibration samples are selected, care should be taken to ensure that all major factors affecting the accuracy of calibration are covered within the limits of the defined application area. These factors include the following:

a) combinations and composition ranges of major and minor sample components to include:

1) range of analytes (e.g. total solids, fat and protein) and non-analytes; or

2) recipes with dairy and non-dairy ingredients and additives (herbs, nuts, spices, chocolate, etc.);

b) seasonal, geographic and genetic effects on milk composition and other raw materials;

c) processing techniques and conditions;

d) ripening stages of cheeses;

e) storage and storage conditions;

f) sample and instrument temperatures and changes in them;

g) instrument variations.

The accuracy of calibrations is influenced by the extent of variation in the sample material and the analyte concentration range. A moderate variation can be easier to fit than a large variation but this depends on the analyte and sample matrices. If the required accuracy cannot be obtained by a single calibration, then stepwise evaluation is an option or the calibration data set should be split up into static or dynamic sub-areas, each with an associated calibration.

It is generally preferable that the whole calibration range is covered in a uniform way, with samples from low to high concentrations of analytes. The sample spread should also be as uniform as possible with respect to the other variables, including those mentioned above. In cases where there are gaps in some areas of the calibration range, all possible actions should be taken to fill these areas, e.g. mixing of samples or using samples from beginning or ending of production. Furthermore, the samples should be collected and measured over a certain period of time to ensure the inclusion of time-dependent effects. This design will improve the ruggedness and give a more even performance of the calibration over the entire analyte concentration range.

Multivariate methods[2][3] may be used as a tool in the selection of samples before the calibration process to ensure a homogeneous calibration set covering all variation in spectroscopic data induced by chemical, biological and physical factors without duplication of samples with similar information. In practice, a larger sample population is measured by NIR spectroscopy for collection of NIR data only. Then samples differing in spectral information are selected for reference analyses. Identification of differing samples may be obtained from inspection of score plots from principal component analysis (PCA) using, for example, the first three components. This may be less practical in the case of many samples. However, it is recommended always to perform a PCA and inspect score plots to obtain a visual overview of the sample set. It is important to check which spectral variations are influencing the loadings and therefore the PCA score plots, i.e. high variation in water or fat content can dominate the PCA results whereas protein variation could be lower but is of importance as well. More formal cluster analyses may be obtained using techniques based on distance measurements[3]. Further samples may be added over a period of time to this pool of selected samples using PCA score space or distance measurement to identify differing samples.

7.2 Reference analyses and NIR instruments

Internationally accepted reference methods for the determination of analytes should be used. The reference method used for calibration should be in statistical control, i.e. the variability should consist of a constant system of random variations. To support the assessment of outliers, it may be useful to perform replicate analyses in independent series (different analysts, different equipment, etc.).

All major variations in NIR measuring conditions that may appear in practice should be considered and built into the calibration model. An important factor is sample temperature, especially for transmittance measurements of aqueous liquids such as milk, etc.

The sampling procedure and sample presentation at the instrument as well as the actual test sample measured by NIR spectroscopy may be critical for the accuracy obtained[4]. The test sample volume or surface interacting in measurements should be large enough to avoid sample inhomogeneity having a significant influence. Reflectance measurement at higher wavelengths normally requires a larger sample surface than transmittance measurement at shorter wavelengths because the light penetration is much less. If possible with the NIR instrument, the optimal sample size should be determined from experiments where the prepared sample material (see 10.1) is measured repeatedly after the refilling of the sample container. Rotating of solid and semi-solid samples helps in acquiring an averaged NIR result for reflectance and diffuse transmittance.

The NIR measurements and reference analyses should preferably be performed on the same sample in order to eliminate effects related to sampling uncertainty. The sample can be divided for NIR measurements and the reference analyses, which should also be performed with a minimum time lag (preferably less than one day). It is good practice to randomize the order in which the samples are presented for both the reference analysis and NIR measurement.

The chemical and physical effects used for the analysis of a certain parameter shall be considered for the reference method and the corresponding NIR method which is calibrated by the reference values. The reference and NIR results may have a bias that is due to differing measurement principles.

7.3 Calibration

Because NIR instruments are applied utilizing different calibration systems, no specific procedure can be given for calibration. However, the person performing the calibration should be familiar with the statistical principles behind the calibration algorithm used.

There are commercially available calibration packages delivered with instruments as part of turn-key solutions. The same principles as described need to be considered when applying such calibrations. An initial validation step is required to check if the commercial methods are applicable for the samples analysed. In some cases, a need for a bias adjustment may occur and this will help to adapt to the current needs. Because the samples and the conditions can be specific an adaptation or expansion of calibrations might be required.

The calibration may be performed using different techniques, e.g. multiple linear regression (MLR), multivariate algorithms such as partial least-squares regression (PLS)[2], locally weighted regression (LWR)[5][6][7] or artificial neural networks (ANNs)[7][8]. The latter techniques are recommended if linearity problems between the spectral response and the constituent occur. Typically, at least 120 calibration samples are needed to obtain rugged calibrations with MLR and PLS. When ANNs are used for calibration, a substantially higher number of samples is required to avoid over-fitting of data because ANNs are very flexible functions with many parameters that shall be determined. Three different data sets (calibration, test or training and validation set) are normally required for determining the architecture, fitting the parameters and validating the ANN calibration. The concept of LWR also requires a considerably larger database from which local calibration samples can be selected for the prediction step.

Spectra should normally be preprocessed prior to calibration to remove or reduce the weighting of effects which are not related to the chemical absorption of light. Frequently used pretreatments are multiplicative scatter correction (MSC)[9], standard normal variate (SNV)[10], de-trending[10] and first or second derivatives[3]. The optimal transformation and other pretreatments of spectra (e.g. smoothing) should be determined by trials. Several techniques may often give equivalent results. The optimal techniques should be assessed from cross validation or independent test sets where models are subsequently developed on parts of the data and tested on other parts[11]. The preprocessing step is important to ensure the robustness of a calibration and one aspect to be checked and considered for the validation process.

An important issue is the selection of the optimal number of variables (in MLR) or factors (in multivariate calibrations). If too few variables or factors are used, an under-fitted model is obtained, which means that the model is not covering enough of important and analyte related variability in the data. If too many variables or factors are used, an over-fitted model may be obtained where noise and minor or irrelevant variations are included in the model. Both cases can result in poor predictions on future samples. The optimal number can be determined by plotting a prediction error [root mean square error of prediction (RMSEP), see 8.2, obtained from an independent test set, or root mean square error of cross validation (RMSECV), see 8.3] versus the number of variables or factors (see Figure B.1). Typically, RMSEP or RMSECV are high for small numbers of variables or factors and decreases as the number increases, before it increases again when the number becomes too high. Generally, the best solution is the one giving the lowest RMSEP or RMSECV with the fewest variables or factors.

The predicted values obtained by cross or test set validation should be plotted against reference results (see Figure B.2). The plot should be examined for outliers and also be investigated for regions with different levels of prediction accuracy, random or systematic, which may indicate the need for more calibration samples or a segmentation of the calibration range.

7.4 Outliers in calibration

7.4.1 General

Outliers may be related to NIR spectra (*x*-outliers) or errors in reference data or samples with a different relationship between reference data and NIR spectra (*y*-outliers).

7.4.2 *x*-outliers

A homogeneous calibration set of spectrally similar samples is required for a robust predictive model. This can also form the basis of an outlier warning system. Any *x*-outliers should thus be removed before calibration. The scores of, for example, the five first PCA or PLS factors can be useful to reveal *x*-outliers either globally outside the population or falling in a gap in the score spaces. A more formal identification of outliers may be performed using, for example, the principle of Mahalanobis distance applied on PCA or PLS reduced data[2][12] or the so-called leverage[13].

It is important to check whether such *x*-outlier candidates are due to extreme high or low concentrations or other reasons which can be related to their identification as outliers. In such a case, the sample is probably important and should be added to the calibration set to widen the range, thus incorporating a different process situation, other batches of ingredients or a changed matrix. In cases where a deviating spectrum is attributed to an insufficient sample preparation or presentation, it shall be removed.

Figure B.2 shows an example plot of NIR prediction versus reference without outliers. In the PCA scores plot in Figure B.3, an *x*-outlier is present.

7.4.3 *y*-outliers

When a *y*-outlier is observed in the calibration set, the reference data should be checked for errors in sample identification, reference analyses, computations, data transfer, etc. However, it may be difficult to relate outliers to errors in reference analyses because the calibration step usually is performed at a later stage than reference analyses, which may make it impossible to repeat analyses because of sample instability. There is no standard way to treat *y*-outliers, but outliers should generally be removed if the difference between NIR predictions and reference results in cross or test set validation exceeds three times the RMSECV or RMSEP (see Clause 8).

It is important to note that the removal of outliers can influence the future prediction of similar samples. Outliers should be removed as a batch before a new calibration model is created. The outlier removal step should only be performed one or two times in order not to reduce the coverage and robustness of the calibration and overestimate the accuracy. Care shall be taken to preserve the optimum distribution of the calibration set when outliers are removed.

Figure B.2 shows an example plot of NIR prediction versus reference without outliers. In Figure B.4, a *y*-outlier is present.

7.4.4 Combined *x*- and *y*-outliers

Samples which are both *x*- and *y*-outliers (influential outliers) have a very strong effect on the regression equation and can be very harmful. Such outliers may give slope effects and increase the prediction error considerably.

Figure B.2 shows an example plot of NIR prediction versus reference without outliers. The outlier in Figures B.4 is a *y*-outlier as the NIR predicted result is different from the reference value due to erroneous reference data or a different relationship between reference data and spectral data. The deviation in prediction may even be caused by an *x*-outlier or spectral outlier which shall be checked and identified by a calibration algorithm specific parameter (e.g. Mahalanobis distance, spectral residuals[2]).

7.5 Validation of calibration models

When calibration equations have been developed, they should be validated on an independent validation set, preferably sampled after the calibration period. The independent validation set should cover all variations in the sample population and should contain at least 25 samples. Only an independent validation set can provide the correct information about accuracy and robustness of an NIR calibration.

The use of cross validation in the calibration process, where subsequent parts of the calibration set are reserved for validation, can give a good estimate of the uncertainty of the method when the calibration samples are properly selected. However, the potential risk is that cross validation may underestimate the ruggedness of the calibration and the predicted uncertainty because cross validation samples are taken from the pool of samples used for calibration. The same applies in the case of using a test set derived from a given calibration set.

The results obtained on the independent validation set are plotted, NIR against reference (or vice versa) such as in Figure B.2 and residuals against reference, to give a visual impression of the performance of the calibration. The standard error of prediction (SEP), see 8.1, is calculated and the residual plot of data corrected for mean systematic error (bias) is examined for outliers, i.e. samples with a residual exceeding 3 × SEP. If an outlier occurs and this cannot be classified as an *x*-outlier and re-analysis of the sample by NIR and reference methods confirms the result, the outlier should not be removed.

In this case, the ruggedness of the calibration is not sufficient and the calibration set should be expanded. The next step is to fit NIR and reference data by linear regression (reference = b × NIR + a) to support the visual impression. If the slope (b) is significantly different from 1, the calibration is skewed. Adjusting the slope of the calibration is generally not recommended. If a re-investigation of the calibration does not detect outliers, especially influential outliers, it is preferable to expand the calibration set to include more samples. However, if the slope is adjusted, the calibration should be tested on a new independent test set. In such cases, more efforts should be put on the validation in terms of number of samples and/or time interval.

The data are also examined for a bias between the methods. An intercept (a) significantly different from 0 indicates that the calibration is biased. A bias may be removed by adjusting the constant term in the calibration equation. However, if the accuracy of the bias-adjusted calibration is significantly poorer than expected from cross validation on the calibration set, i.e. SEP is significantly larger than RMSECV, the calibration set should be expanded to include more samples.

In all cases when a new calibration is developed on an expanded calibration set, the validation process should be repeated on a new independent test set. If necessary, expansion of the calibration set should be repeated until acceptable results are obtained on an independent test set.

7.6 Changes in measuring and instrument conditions

If measurement or instrument conditions are changed, the stated accuracy of a validated NIR method cannot generally be considered as valid unless an additional validation is performed.

For example, the calibrations developed for a certain population of samples may not be valid for samples outside this population, although the analyte concentration range is unchanged. A calibration developed on cheeses from one dairy may not give the same accuracy on cheeses produced in another dairy if the processing and ripening parameters are different.

Changes in the sample preparation and presentation technique or the measuring conditions (e.g. temperature) not included in the calibration set may also influence the analytical results. If the conditions are changed, a supplementary validation should be performed.

Furthermore, calibrations developed on a certain instrument cannot always be transferred directly to an identical instrument operating under the same principle. It may be necessary to perform bias adjustments to calibration equations. In some cases, it may even be necessary to standardize the two instruments against each other by mathematical procedures before calibration equations can be transferred[3]. Standardization procedures may be used to transfer calibrations between instruments of different types provided that samples are measured in the same way (reflectance, transmittance) in similar containers and that most of the spectral region is common. Adding a few samples scanned with the second instrument in the database can contribute to the transfer.

The calibrations should be checked and validated whenever any major part of the instrument (optical system, detector) has been changed or repaired.

7.7 Outlier detection

Use of NIR methods is generally limited to samples in the population covered by the calibration set with respect to sample material characteristics and analyte concentrations. An outlier detection system should accompany the NIR method to reduce the risk of unintentional use of NIR spectrometry on samples outside this population. The system should be able to detect *x*-outliers and samples falling outside the concentration range. The principle of Mahalanobis distance applied to PCA or PLS scores or “leverage” may be used for the detection of *x*-outliers. If a sample is detected as an outlier, the sample should be re-analysed by reference methods to obtain the best estimate. In cases where the NIR result is not a *y*-outlier, the sample spectrum can be used for a calibration update. In cases where it is a *y*-outlier, the calibration should be updated if such samples are occurring frequently within routine operation.

8 Statistics for performance measurement

8.1 Standard error of prediction (SEP) and bias

The key statistical parameter for the performance of a calibration[14] is the SEP. It expresses the accuracy of routine results corrected for the mean difference between routine and reference methods (bias) and is calculated by using the Formula (1):

$$\mathrm{SEP}=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(y_{\mathrm{NIR},i}-y_{\mathrm{Ref},i}-B\right)^2} \qquad (1)$$

where

$\left(y_{\mathrm{NIR},i}-y_{\mathrm{Ref},i}\right)$ is the difference between results obtained by the NIR method $y_{\mathrm{NIR},i}$ and the reference method $y_{\mathrm{Ref},i}$ on sample *i*;

B is the bias, explained by the formula:

$$B=\frac{1}{n}\sum\left(y_{\mathrm{NIR},i}-y_{\mathrm{Ref},i}\right)=\bar{y}_{\mathrm{NIR}}-\bar{y}_{\mathrm{Ref}};$$

n is the total number of samples.

SEP is equal to the standard deviation of predicted residuals. This term expresses the accuracy which can be expected with the calibration equation when the bias is zero.

To obtain realistic estimates of the accuracy, SEP should be calculated on samples of an independent validation set outside the calibration set and the distribution of the analyte concentration range should be as uniform as possible.

When implementing a bias, it is important to ensure that the value used truly reflects a significant systematic difference between the predicted result and reference values. It is important to assess multiple results over time alongside an assessment of result significance by means of classic statistic methods to ensure that a bias adjustment is required and justified. It should not be applied or implemented based on a few short-term or erroneous results.

8.2 Root mean square error of prediction (RMSEP)

Instead of reporting SEP and bias in separate terms, they may be included in a single term, the RMSEP, defined as shown in Formula (2):

$$\mathrm{RMSEP}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_{\mathrm{NIR},i}-y_{\mathrm{Ref},i}\right)^2} \quad (2)$$

When the difference between the NIR method and the reference method is not clearly systematic, the SEP may overestimate the possibility of improving the accuracy using bias adjustment. In this case, the RMSEP gives a more realistic estimate of the prediction capability of the calibration. When the bias is insignificant, the RMSEP tends towards SEP with increasing number of samples.

To obtain realistic estimates of the accuracy, RMSEP should be calculated on samples of an independent validation set outside the calibration set and the distribution of the analyte concentration range should be as uniform as possible.

8.3 Root mean square error of cross validation (RMSECV)

Formula (2) is also used for RMSECV. The difference is that the RMSECV is calculated from cross validation predictions on the calibration set and not on an independent validation set.

9 Sampling

Sampling is not part of the method specified in this document. A recommended sampling method is given in ISO 707 | IDF 50[15].

It is important that the laboratory receives a sample that is truly representative and has not been damaged or changed during transport or storage. All laboratory samples should be kept at conditions suitable for the product and the production process from the time of sampling to the time of commencing the procedure. A sample, for example, could be cooled, heated to room temperature or stored under specific conditions to preserve its representativeness.

10 Procedure

10.1 Preparation of test samples

10.1.1 General aspects

Keep the prepared sample in an airtight container until the measurement, which preferably should be carried out within 24 h. If delay is unavoidable, take every precaution to ensure proper storage of the sample. When refrigerated, ensure that any moisture condensation on the inside surface of the container is thoroughly and uniformly re-incorporated into the test sample.

Mass reduction of the prepared sample to obtain the analytical sample should be performed by principles that keep the sampling error to a minimum. Use of incremental sampling techniques (e.g. riffle splitters for powdered products) may be useful for heterogeneous materials.

10.1.2 Milk and liquid milk products

Samples in liquid form are, for example, milk (raw, skim, processed, condensed), whey, cream, concentrates, milk drinks, premixes. For a transmission measurement, they can be filled into a flow cell with, for example, 1 mm path length either with a high-pressure pump with a homogenizer valve, by a peristaltic pump or in another appropriate way. For transflectance measurements, a suitable setup and path length shall be used in terms of sample container and reflector/mirror.

Some sample types such as raw milk and cream show separation of fat or creaming. These sample types should therefore be mixed directly before the measurement by gently inverting the sample container 5 to 10 times after pre-warming to 40 °C.

10.1.3 Cheese

Before the analysis, remove the coating and/or rind from the cheese in such a way as to obtain a sample representative of the cheese as it is usually consumed. Prepare the sample using an appropriate device (see 6.2). Quickly mix the ground or grated sample. If the sample cannot be ground or grated, mix it thoroughly by intensive kneading. Care should always be taken to avoid moisture loss.

With different instruments and techniques, the preparation of the cheese sample can have an impact.

— For diffuse transmission, a reproducible thickness of the sample layer is mandatory. However, the results can be influenced by the density of the sample related to the process of filling and compacting by the operator as well as by the cheese type.

— Diffuse reflectance top-down measurements can by influenced by the drying of the top layers of sample and this may be a possible error source. Similarly, for finely cut cheeses, some degree of segregation may have an effect for measurements from below due to the segregation of fine particles and thus different light scattering properties of the sample.

10.1.4 Dried milk, whey, buttermilk and whey protein powders

Thoroughly mix the sample by repeatedly rotating and inverting the sample container (if necessary, after having transferred the entire laboratory sample to an airtight container of sufficient capacity to allow this operation to be carried out). In cases of a risk of segregation, the sample is transferred to the container without any further steps. Just pouring the sample into the container has again a risk of segregation and should be avoided, e.g. by using a spoon or spatula-like scoop.

10.1.5 Butter

Mixing is not necessary unless the sample gives evidence for it. In that case, the mixing temperature should not exceed 25 °C.

10.1.6 Yoghurt

Mixing is not necessary unless the sample gives evidence for it. Fruit pieces or other added particles may be removed according to the local quality control procedure.

10.2 Measurement

The prepared sample should reach a temperature appropriate for the sample type and within the range included in the calibration, e.g. 15 °C to 25 °C for cheese and milk powder, and 8 °C to 12 °C for butter. A sub-sample is transferred to:

— a cell or cuvette for liquids;

— the transflectance setup for liquids;

— a sample container or, for example, Petri dish for solid and semi-solid without compressing. In cases of a measurement from the top, the sample surface should be levelled off with a minimum of disturbance.

The sample is measured following the instructions given by the manufacturer for the operation of the instrument.

The number of scans, or sub-sample predictions averaged, or the dwell time on each wavelength should be large enough to reduce the noise of the measurement to an insignificant level for the given application and the manufacturer guidelines. The calibration model that is valid for the measured sample type is then applied. If re-usable sample containers with windows are used, the windows should be cleaned between measurements, i.e. particles may be removed by a brush or a vacuum cleaner and sticky compounds may, for example, be removed using a cloth moistened with a suitable solvent (see Clause 5).

10.3 Evaluation of results

Results obtained on samples detected as spectral or concentration outliers may not be regarded as reliable.

11 Checking instrument stability

11.1 Control samples

A control sample is a material either from a reference material provider or selected from the sample types routinely measured by the instrument, and is assessed at least once daily. These samples are used for testing the instrument and procedure stability, by analysing them and checking the performance of sample preparation, sample presentation, instrument and prediction model. A requirement is the selection of an as-stable-as-possible control sample and an estimation of how long it can be reliably used for.

Knowledge of the true concentration of the analyte in the control sample is not necessary. The sample material should, as far as possible, resemble the samples to be analysed. The parameter measured should be identical to, or at least biochemically similar to, the sample analyte. One type of control sample per measurement setup is sufficient to test this via the reflectance or transmittance modules.

Liquid milk products, cheese, milk powder and butter may be used as pilot samples. A check sample is prepared as in 10.1 and may be stored refrigerated as a series of sub-samples in closed containers. These samples should be stable for at least one week. The check sample stability should be tested in actual real-life situations. Shifts between control samples should have an overlap to secure uninterrupted control.

The recorded day-to-day variation should be plotted in control charts and investigated for significant patterns or trends over time.

11.2 Instrument diagnostics

For scanning spectrophotometers, the wavelength/wavenumber accuracy and precision and the photometric instrument noise should be checked at least once per week, or more frequently if recommended by the instrument manufacturer, and the results should be compared to specifications and requirements. Daily tests can indicate timely warnings of lamp failure, mechanical instrument problems, light leaks in the sampling area, excessive temperature and humidity variations for example. These tests may be part of a self-executing diagnostic system built into the instrument.

12 Running performance check of calibration

NIR methods should be validated continuously against reference methods to secure steady optimal performance of calibrations and observance of accuracy. The frequency of checking the NIR method should be sufficient to ensure that the method is operating under steady control with respect to systematic and random deviations from the reference method. The frequency depends on, among other things, the application, the process and its variation, accuracy needs, laboratory capacity and, typically, costs. The number of samples analysed per day must be defined by the needs and the strategy of the quality control group of the company.

The routine validation should be performed on samples selected randomly from the pool of analysed samples. It may be necessary to resort to some sampling strategy to ensure a balanced sample distribution over the entire calibration range, e.g. segmentation of concentration range and random selection of independent test samples within each segment, product and/or recipe type.

Results should be assessed by control charts, plotting running sample numbers on the abscissa and the difference between results obtained by reference and NIR methods on the ordinate; ±2 × SEP (95,5 % confidence) and ±3 × SEP (99,7 % confidence) may be used as warning and action limits where the SEP has been obtained on a validation set collected independently of calibration samples.

Control charts should be checked for systematic bias drifts from zero, systematic patterns and excessive variation of results. General rules applied for Shewhart control charts may be used in the assessment[16]. However, too many rules applied simultaneously may result in too many false alarms.

The following rules used in combination have proved to be useful in detection of problems:

— one point outside either action limit;

— two out of three points in a row outside a warning limit;

— nine points in a row on the same side of the zero line.

Additional control charts plotting other features of the running control (e.g. mean difference between NIR and reference results, see ISO 9622 | IDF 141[17]) and additional rules may be applied to strengthen decisions.

To reduce the risk of false alarms, the control samples should be analysed independently (in different series) by both NIR spectrometry and reference methods to avoid the influence of day-to-day systematic differences in, for example, reference analyses.

If the warning limits are often exceeded and the control chart only shows random fluctuations (as opposed to trends or systematic bias), the control limits may have been based on a too optimistic SEP. An attempt to force the results within the limits by frequent adjustments of the calibration will not improve the situation in practice. The SEP should instead be re-evaluated using the latest results.

If, after a period of stability, the calibration equations begin to move out of control, the calibration should be upgraded. Before this is done, an evaluation should be made of whether the changes could be due to changes in the raw materials, the production process, recipes, reference analyses, unintended changes in measuring conditions (e.g. caused by a new operator), instrument drift or malfunction, etc. In some cases, a simple adjustment of the constant term in the calibration equation may be sufficient (an example is shown in Figure B.5). In other cases, it may be necessary to run a complete re-calibration procedure, where the complete or a part of the basic calibration set is expanded to include samples from the running validation, and perhaps additional samples selected for this purpose (an example is shown in Figure B.6).

Considering that the reference analyses are in statistical control and the measuring conditions and instrument performance are unchanged, significant biases or increased SEP values can be due to changes in the chemical, biological or physical properties of the samples compared to the underlying calibration set. Such changes could in practice be caused by, for example, changes in the cheese processing parameters.

13 Precision and accuracy

13.1 Repeatability

The repeatability, i.e. the difference between two individual single test results obtained with the same method on identical test material in the same laboratory by the same operator using the same equipment within a short interval of time, should not be exceeded in more than 5 % of cases. This depends on the sample material, the analyte, sample and analyte variation ranges, method of sample presentation, instrument type, and the calibration strategy used. The repeatability should be determined in each use case.

13.2 Intra-laboratory reproducibility

The intra-laboratory reproducibility, i.e. the difference between two individual single test results obtained on identical test material in the same laboratory by different operators at different times, should not be exceeded in more than 5 % of cases. This depends on the sample material, the analyte, sample and analyte variation ranges, method of sample presentation, instrument type, and the calibration strategy used. The intra-laboratory reproducibility should be determined in each use case.

13.3 Accuracy

The accuracy, which includes uncertainty from systematic deviation from the true value on the individual sample (trueness) and uncertainty from random variation (precision), depends *inter alia* on the sample material, the analyte, sample and analyte variation ranges, method of sample presentation, instrument type, and the calibration strategy used. The accuracy should be determined in each use case. SEP and RMSEP values reported in the literature are listed in Table A.1. The reported SEP and RMSEP values also include uncertainty of reference results, which may vary from case to case. More calibration and validation results are published in an IDF bulletin[1].

14 Test report

The test report shall specify:

a) all information necessary for complete identification of the sample;

b) the test method used, with reference to this document, i.e. ISO 21543 | IDF 201;

c) all operating conditions not specified in this document, or regarded as optional,

d) any circumstances that could have influenced the results;

e) the test result(s) obtained;

f) the current RMSEP and bias (if statistically significant), estimated from running a performance test on at least 25 validation samples (see Clause 12);

g) the date of the test.

Annex A
(informative)

Examples of SEP and RMSEP values

The SEP and RMSEP values given in Table A.1 have been reported in literature. The reported SEP and RMSEP values also include uncertainty of reference results, which may vary from case to case.

Table A.1 — SEP and RMSEP values

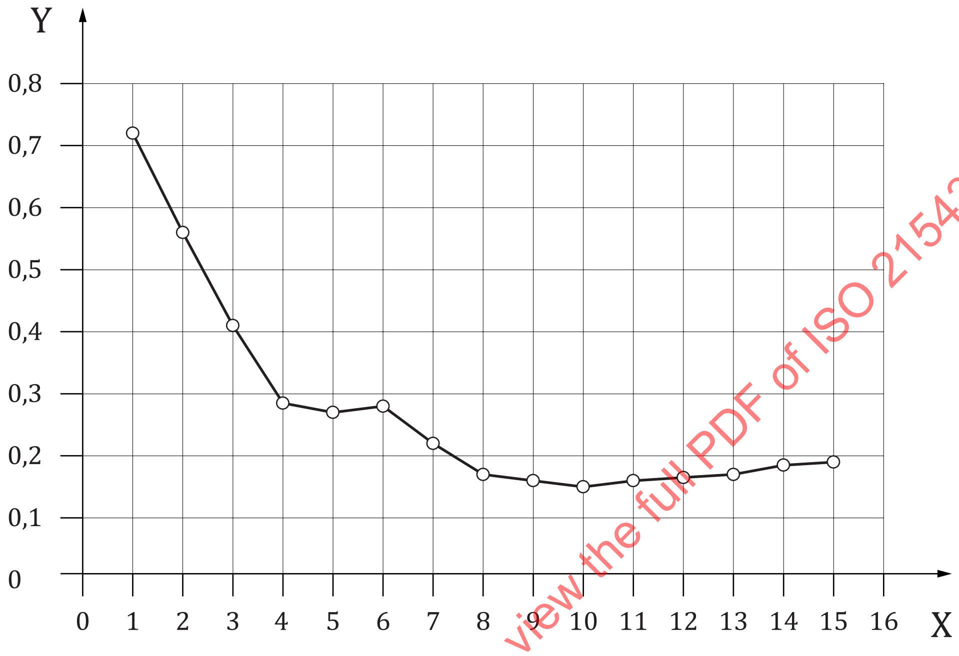
Sample material	Analyte	Conc, range %	RMSEP %	SEP %	NIR technique	Ref.
Non-homogenized milk samples	Fat	0,03 to 6,22		0,05	Transmittance	[18]
	Crude protein	2,03 to 3,32		0,09		
	Total solids	1,88 to 3,16		0,12		
	Casein	1,56 to 2,69		0,07		
Non-homogenized raw milk	Fat	2,63 to 5,47		0,03	Transmittance	[19]
	Protein	2,86 to 3,81		0,07		
	Lactose	3,99 to 4,80		0,09		
	Total solids	11,27 to 14,63		0,12		
	Non-fat solids	8,21 to 9,49		0,11		
Non-homogenized raw milk	Fat	3,29 to 6,84	0,05		Reflectance	[20]
	Protein	2,80 to 4,68	0,11		Transmittance	
	Lactose	3,48 to 5,09	0,18			
	Fat	3,29 to 6,84	0,04			
	Protein	2,80 to 4,68	0,13			
	Lactose	3,48 to 5,09	0,13			
Non-homogenized raw milk	Fat	0,7 to 12,3		0,09	Reflectance	[21]
	Protein	2,4 to 4,0		0,05		
	Lactose	3,9 to 5,2		0,06		
Non-homogenized raw milk	Fat	1,9 to 7,4		0,04	Reflectance	[22]
	Protein	2,6 to 4,1		0,08		
	Lactose	4,4 to 5,3		0,09		
Processed cheese, Gouda, Edam	Moisture	40 to 51		0,24	Reflection	[23]
	Fat	21 to 31		0,27	Ground samples	
Processed cheese	Moisture	48 to 51		0,21	Reflection	[23]
	Fat	21 to 23		0,23		
	Protein	20 to 24		0,35		
Cheddar	Moisture	35 to 40	0,34		Reflection	[24]
	Fat	31 to 35	0,33		Ground samples	
Tetilla, Arzúa, Edam	Total solids	45 to 62	0,61		Reflection	[25]
	Fat	18 to 32	0,47		Unground samples	
	Protein	16 to 30	0,50			

Table A.1 *(continued)*

Sample material	Analyte	Conc, range %	RMSEP %	SEP %	NIR technique	Ref.
Danbo	Moisture Fat Protein	40 to 52 22 to 28 22 to 27	0,30 0,28 0,26		Transmission Unground samples	[26]
Danbo	Total solids Fat Protein	46 to 62 14 to 36 22 to 31	0,58 0,52 0,38		Reflection Unground samples	[27]
Edam	Total solids	50 to 61	0,20		Transmission Ground samples	[28]
Gouda	Total solids	40 to 43	0,12		Transmission	[28]
Brie	Total solids	41 to 55	0,33		Transmission	[28]
Colby	Total solids	38 to 41	0,23 to 0,27		Transmission	[28]
Cheddar	Total solids	37 to 40	0,31 to 0,35		Transmission	[28]
Danbo	Total solids Fat	50 to 63 23 to 29	0,20 0,19		Transmission Unground samples	[29]
Danbo	Total solids Total solids Fat	47 to 52 47 to 63 16 to 28	0,16 0,29 0,17		Transmission Unground samples	[30]
Dried skim milk	Moisture Fat Protein Lactose	3,3 to 4,7 0,5 to 1,3 34 to 37 48 to 50		0,08 0,09 0,20 0,44	Reflection	[23]
Dried buttermilk	Moisture Fat Protein Lactose	2,7 to 6,3 5,3 to 11 29 to 35 37 to 47		0,10 0,13 0,21 0,37	Reflection	[23]
Dried whole milk	Moisture Fat Protein	2,0 to 4,1 24,3 to 29,6 22,2 to 28,5		0,08 0,15 0,13	Reflection	[31]
Dried skim milk	Moisture Fat Protein Lactose	2,9 to 9,7 0,5 to 2,1 34 to 40 53 to 58		0,27 0,10 0,44 0,59	Reflection	[32]
Dried whey	Moisture Fat Protein Lactose	2,7 to 5,7 0,2 to 7,2 9,5 to 42 7,9 to 71		0,37 0,52 1,3 2,8	Reflection	[33]
Butter	Moisture Non-fat solids Fat	14 to 16 1,3 to 2,6 81 to 84	0,26 0,071 0,38		Reflection	[34]

Annex B
(informative)

Examples of figures



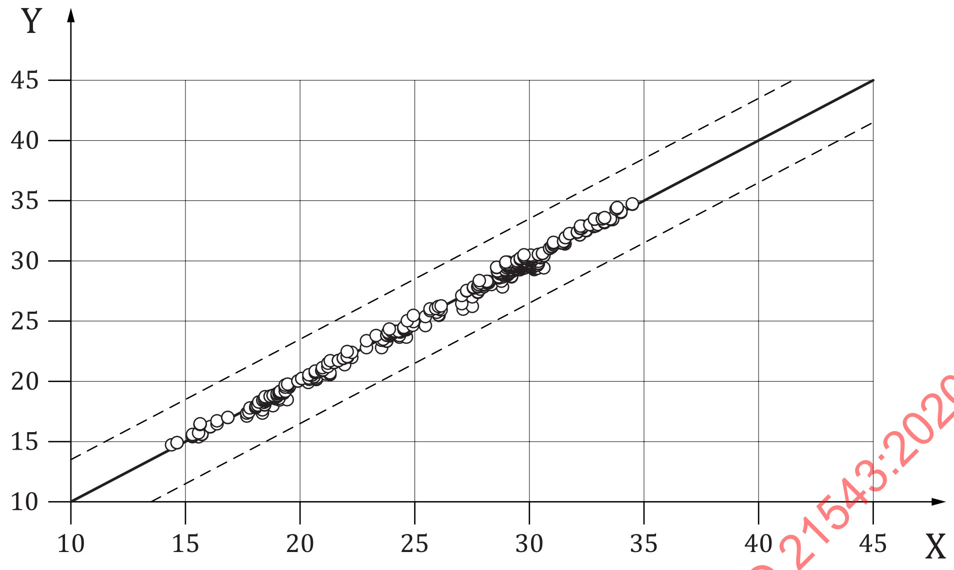
Key

X number of PLS factors

Y RMSECV or RMSEP

NOTE The optimal number of PLS factors is 10 ± 1.

Figure B.1 — Example plot of RMSECV or RMSEP as function of the number of PLS factors

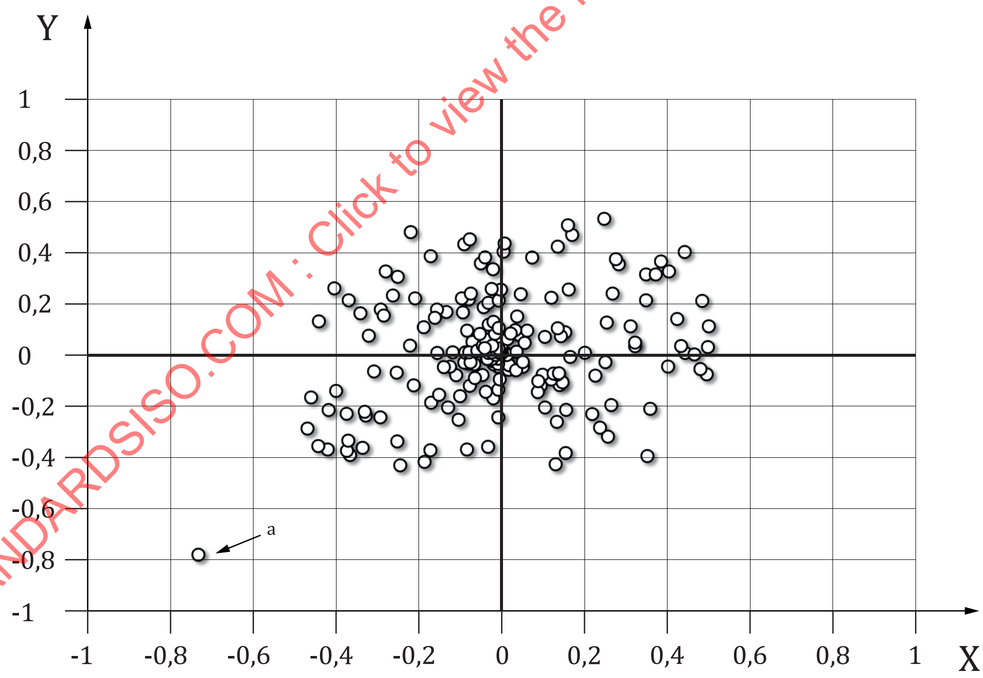


Key

X reference value

Y NIR value

Figure B.2 — Example of NIR prediction versus reference plot with no outliers



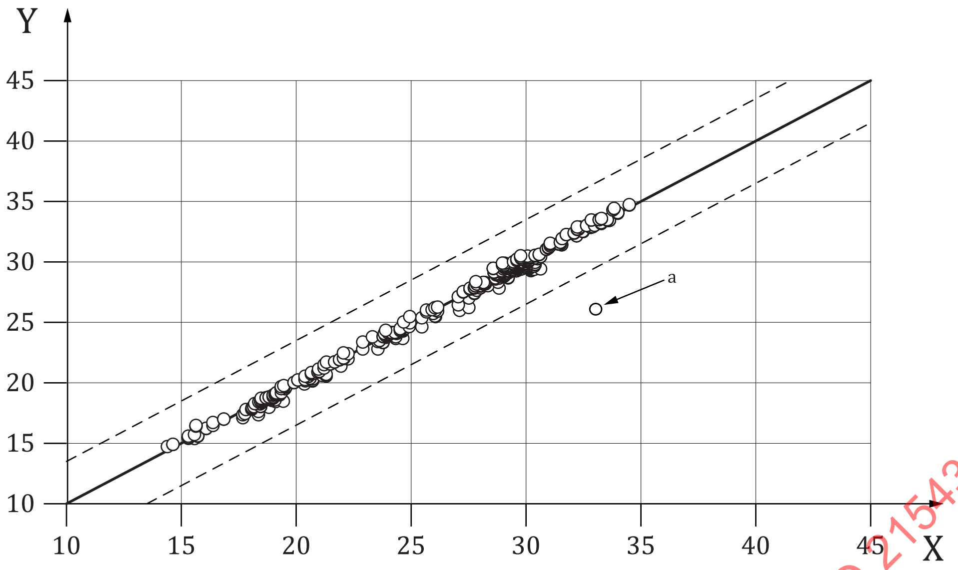
Key

X scores factor 1

Y scores factor 2

a Outlier.

Figure B.3 — Example of PCA score plot with an *x*-outlier



Key

X reference value

Y NIR value

[a] Outlier.

Figure B.4 — Example of NIR prediction versus reference plot with *y*-outlier

The plot of predicted versus reference values (or vice versa) in Figure B.4 shows one sample that strongly deviates from the other samples. If the reason for this deviation is not the NIR spectrum (*x*-outlier), this sample is a *y*-outlier, due to erroneous reference data or a different relationship between reference data and spectral data.

In Figure B.5, no points are outside the upper action limit (UAL) or the lower action limit (LAL). However, nine points in a row (e.g. 14 to 22) are on the same side of the zero line. That indicates a bias problem and in addition two points (27 and 28) out of three points are outside the lower warning limit (LWL) but none is outside the upper warning limit (UWL).