Talanta 85 (2011) 562–568

Contents lists available at ScienceDirect

Talanta

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

# Melamine detection by mid- and near-infrared (MIR/NIR) spectroscopy: A quick and sensitive method for dairy products analysis including liquid milk, infant formula, and milk powder

Roman M. Balabin[a,*], Sergey V. Smirnov[b]

[a] Department of Chemistry and Applied Biosciences, ETH Zurich, 8093 Zurich, Switzerland
[b] Unimilk Joint Stock Co., 143421 Moscow region, Russia

ARTICLE INFO

*Article history:*
Received 5 January 2011
Received in revised form 7 April 2011
Accepted 11 April 2011
Available online 19 April 2011

*Keywords:*
Food (milk-derived products)
Dairy/milk products (liquid milk, infant formula, milk powder)
Partial least squares regression (PLS)
Artificial neural networks (ANN)
Support vector machines (SVMs: SVR and LS-SVM)
Biofuel (ethanol–gasoline fuel, vegetable oil, biodiesel)

ABSTRACT

Melamine (2,4,6-triamino-1,3,5-triazine) is a nitrogen-rich chemical implicated in the pet and human food recalls and in the global food safety scares involving milk products. Due to the serious health concerns associated with melamine consumption and the extensive scope of affected products, rapid and sensitive methods to detect melamine's presence are essential. We propose the use of spectroscopy data-produced by near-infrared (near-IR/NIR) and mid-infrared (mid-IR/MIR) spectroscopies, in particular—for melamine detection in complex dairy matrixes. None of the up-to-date reported IR-based methods for melamine detection has unambiguously shown its wide applicability to different dairy products as well as limit of detection (LOD) below 1 ppm on independent sample set. It was found that infrared spectroscopy is an effective tool to detect melamine in dairy products, such as infant formula, milk powder, or liquid milk. ALOD below 1 ppm ($0.76 \pm 0.11$ ppm) can be reached if a correct spectrum preprocessing (pretreatment) technique and a correct multivariate (MDA) algorithm—partial least squares regression (PLS), polynomial PLS (Poly-PLS), artificial neural network (ANN), support vector regression (SVR), or least squares support vector machine (LS-SVM)—are used for spectrum analysis. The relationship between MIR/NIR spectrum of milk products and melamine content is nonlinear. Thus, nonlinear regression methods are needed to correctly predict the triazine-derivative content of milk products. It can be concluded that mid- and near-infrared spectroscopy can be regarded as a quick, sensitive, robust, and low-cost method for liquid milk, infant formula, and milk powder analysis.



## 1. Introduction

Melamine or 2,4,6-triamino-1,3,5-triazine is a nitrogen-rich chemical implicated in the pet and human food recalls in 2007 [1,2] and in the global food safety scares in 2008 involving milk products [3–5]. In those food safety incidents, melamine was intentionally added to foods and animal feed to boost the protein content [6,7]. A driving force for the adulteration of a food product with melamine is that it has high nitrogen content, which increases the apparent protein content measured by standard protein analysis tests, such as the Kjeldahl or Dumas methods [5]. Note that the Kjeldahl analytical method is a method only for the quantitative determination of nitrogen (not true protein content) [8].

In late 2008, trace amount of melamine were detected in US-made infant formula products [8–10]. The recalls involving pet food and milk products, contaminated with melamine, have created a widespread food safety scare [5,8]. Today melamine contamination has been reported in a variety of food products, such as milk (liquid or powder), infant formula, frozen yoghurt, pet food, biscuits, candy, and coffee drinks [11–15]. See Ref. [8] and references therein for a good introduction and a problem setting.

Current analytical methods for testing melamine are mainly chromatography-based methods (e.g., HPLC, GC–MS), which are time consuming, expensive, and labor-intensive and require complex procedures of sample pretreatment and well-trained technicians to operate the instrumentation [1]. Therefore, there is an urgent and increasing need in the food industry and analytical chemistry to develop simpler, quicker, and cost-effective methods for detecting melamine and its analogues (e.g., cyanuric acid, melamine cyanurate, ammelide, and ammeline) in food ingredients and processed foods and to develop reliable predictive models for quantifying melamine in foods [1].

The ideal characteristics of an analytical technique for detecting melamine include the following: (1) high sensitivity or low limit of detection (LOD), below 1 ppm; (2) high specificity; (3) short

* Corresponding author. Tel.: +41 44 632 4783.
*E-mail addresses:* balabin@org.chem.ethz.ch, balabinrm@yandex.ru (R.M. Balabin).

0039-9140/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.talanta.2011.04.026

detection time; (4) low-cost; (5) portability; (6) minimum or no sample preparation; and (7) suitability for measuring melamine in a variety of complex food matrixes [15–19]. Unfortunately, no existing method meets all of the aforementioned requirements [1,20,21].

We propose the use of spectroscopy data-produced by near-infrared (near-IR or NIR) and medium-infrared (mid-IR or MIR) spectroscopies, in particular—for melamine detection in complex matrixes.

In general, spectral data are components of data from techniques such as near-infrared (NIR) spectroscopy [22,23], mass spectrometry (MS) [24], and nuclear magnetic resonance ($^1$H NMR/$^{13}$C NMR) spectroscopy [25]. Of these, NIR spectroscopy has advantages over other analytical tools because it is noninvasive, requires minimal sample preparation, and can yield a response in real time [21] (an important advantage for processing analytical chemistry, PAC [26]). NIR spectroscopy is based on the absorption of electromagnetic radiation in the region of 780 to 2500 nm (12,820–4000 cm$^{-1}$) [27].

Various analytical studies of NIR spectra have been conducted throughout the past two decades. Applications of NIR spectroscopic data can be found in medical and biomedical studies, food science, forestry, and the pharmaceutical and petroleum industries [22,25] (see Ref. [21] for additional references). Vibrational spectroscopy techniques (IR, NIR, and Raman) [28–31], when associated with multivariate data analysis (MDA), have proven to be powerful tools in the analysis of fuel samples, such as gasoline, diesel, alcohol fuel (ethanol–gasoline mixtures [32,33]), and kerosene (jet fuel) [21]. These spectroscopic methods are much faster than the usual (e.g., LC-/HPLC-based) techniques, present good accuracy and precision, are nondestructive, and can be used in remote quality control [34–37].

Analysis of NIR spectra usually involves a combination of multiple samples, each of which has a large number of correlated features [21,23,35,38,39]. As such, a variety of data mining algorithms have been introduced to reduce the complexity accompanying such large amounts of data, aiming to identify meaningful patterns in NIR spectra. Multivariate calibration methods have been increasingly used to extract relevant information from different types of spectral data to predict analyte concentrations or properties of complex samples [1,38,39]. However, the main problem associated with these methods is the nonlinearity of the data [38,39].

Several strategies have been used for the calibration of nonlinear data systems [38], such as data pretreatment (e.g., data transformation and variable selection), using linear methods (for slight nonlinearities only), local modeling, adding extra variables, and using nonlinear calibration techniques. Among these approaches, nonlinear calibration techniques are the only ones able to build robust calibration models [40]. Such calibration models have potential formodeling severe intrinsic nonlinearities that can be found in natural "sophisticated" multicomponent systems (e.g., biodiesel fuel [27] or milk powder [36]).

The most important linear calibration method is the partial least squares/projection on latent structures (PLS) regression [30,39]. The two most important nonlinear calibration methods are the nonlinear variants of PLS (e.g., polynomial PLS) and artificial neural networks (ANNs) [39]. Several comparative studies on these two techniques have been conducted using various data sets. In some studies, the neural networks performed better than PLS when the data were nonlinear (also see above) [10,38,39,41]. In other situations, ANN and nonlinear PLS gave equally good results [42]. It is possible that the different conclusions obtained from the various studies resulted from differences in the nature of the nonlinearities [38,39].

Near-infrared spectroscopy in combination with MDA methods has been applied to melamine detection in a number of analytical studies.

In 2009, Mauer et al. [8] evaluated near- and mid-infrared spectroscopy methods (NIR, FTIR-ATR, FTIR-DRIFT) for the detection and quantification of melamine in infant formula powder. Partial least squares models were established for correlating spectral data to melamine concentration: $R^2 > 0.99$, RMSECV $\leq 0.9$, and residual prediction deviation (RPD) above 12. Factorization analysis of spectra was able to differentiate unadulterated infant formula powder from samples containing 1 ppm melamine with no misclassifications, a confidence level of 99.99%, and a selectivity >2. It was stated that NIR and MIR methods enable rapid detection of 1 ppm melamine in infant formula powder [8].

Unfortunately, no independent test was conducted using extra sample sets to estimate the real prediction error (RMSEP). It should also be noted that a calibration error of 0.616 ppm (the best value reported [8]) could lead to a LOD value above 1 ppm (LOD = $3 \times E$ = 1.848 ppm). To be applicable in food industry, the method should definitely have a LOD value below 1 ppm. The generality of the IR-based method for the analysis of other dairy products is also not clear since only infant formula samples were tested.

Lu et al. [36] established a novel and rapid method for detecting pure melamine in milk powder using near-infrared spectroscopy based on least squares-support vector machine (LS-SVM). Partial least square discriminant analysis (PLS-DA) was used for the extraction of principal components (PCs). The scores of the first two PCs have been applied as inputs to LS-SVM. Compared to PLS-DA, the performance of LS-SVM was better, having higher classification accuracy—both 100% for the training and testing set. The detection limit was lower than 1 ppm. Based on the results, it was concluded that NIR spectroscopy combined with LS-SVM could be used as a quick and accurate method to detect pure melamine in milk powder.

The generality of the proposed NIR-based method [36] is not clear since only milk powder samples were tested. Note also that only the classification task (not regression) was solved by the NIR + LS-SVM approach. Thus, in this case, the RMSEP is also not reported [36].

To finalize the previous paragraphs, none of the up-to-date reported NIR- or IR-based methods for melamine detection has unambiguously and simultaneously shown its wide applicability to different dairy products as well as LOD below 1 ppm on independent (test) sample set.

In this paper, we have two goals: (i) to establish a quick, sensitive (LOD <1 ppm), reliable, and robust method for melamine detection in different dairy products (liquid milk, infant formula, milk powder) based on NIR and MIR spectroscopy methods; (ii) to compare different multivariate calibration models (e.g., PLS, OPLS, ANN, SVM) on large melamine data sets (>660 samples each) to find the best candidate for industrial analytical application. A large calibration range of melamine concentration (0.11–2000 ppm) was used to build a model that is capable of dealing with both almost melamine free and largely contaminated dairy samples. Notably, the artificial neural networks have never been used to predict melamine content in milk products from vibrational spectral data.

## 2. Experimental

### 2.1. Materials

Six hundred ninety (690) infant formula samples, six-hundred sixty (660) milk powder samples, and six hundred sixty (660) liquid milk samples were prepared for MIR/NIR analysis from standard (noncontaminated) products, supplied by Unimilk Joint Stock Co. and purchased in a local store. Such a huge amount of samples allows for the estimation of the efficiency of the model at the "sample set limit"–see the "basis set limit" (BSL) or "complete basis set"

**Table 1**
Description of sample sets: infant formula, milk powder and liquid milk.

| | | Infant formula | Milk powder | Milk |
|---|---|---|---|---|
| Material class (type) | | Powder | Powder | Liquid |
| Initial number of dairy products | | 60 | 72 | 72 |
| Number of samples | Total | 690 | 660 | 660 |
| | Cal (CV[a]) | 621 (69) | 594 (66) | 594 (66) |
| | Test | 69 | 66 | 66 |
| Calibration property | Name | Melamine content | | |
| | Unit | ppm (mg kg$^{-1}$) | ppm | ppm |
| Property range | Range | 0.15–2000 | 0.11–2000 | 0.11–2000 |
| | Mean ± std | 222 ± 437 | 218 ± 440 | 218 ± 440 |
| Subranges | Border value | 17.3 | 14.6 | 14.6 |
| | [Low] | 0.15–30.2 | 0.11–30.2 | 0.11–30.2 |
| | [High] | 9.87–2000 | 7.02–2000 | 7.02–2000 |
| Reference method accuracy | | 0.1–0.5% | 0.1–0.5% | 0.2–0.5% |

[a] Nine-fold cross-validation (CV) procedure was applied.

(CBS) in quantum chemistry [43–46], e.g. in focal point analysis (FPA) schemes [43,47,48]. The initial samples (60, 72, and 72 for infant formula, milk powder, and liquid milk, respectively) were checked for the absence of melamine contamination using the standard HPLC-based method [49].

The initial 60/72 samples were mixed in random proportions to produce 690/660 different chemical systems (as it was done by one of us in Ref. [22] for petroleum macromolecules). Melamine was added to each sample independently. Such an experimental design ensures the robustness of the final NIR/MIR model to the variations in chemical composition of the dairy products.

Liquid milk samples have protein and fat contents of 2.7–5.8% and 0.5–10.4%, respectively. Their lactose content was between 4.2 and 5.9%. Moisture content in the milk powder was 4–12%. Infant formula contained from 1.02 to 1.30 g oz$^{-1}$ of fat.

Detailed information about the samples sets can be found in Table 1. The range of melamine concentration was set to be from very low (0.11 mg kg$^{-1}$ or 0.11 ppm) to very high (2000 ppm). This range ensures the stability and robustness of the final prediction model independent from the concentration of contaminating substance.

During the relatively long measurement period we have used four (4) different melamine samples from three (3) producers. Dry samples were prepared according to Ref.[36]. From 1 to 5 g of the sample was prepared in each run. Samples were well homogenized before experiments. Experiments were carried out within a very short span after sample preparation in order to minimize experimental errors. Sample homogeneity was checked by NIR and MIR spectroscopy methods.

## 2.2. Methods

### 2.2.1. Near-infrared spectroscopy

The near-infrared spectra were acquired with a MPA Multi Purpose FT-NIR Analyser (Bruker, Germany). The spectra were acquired at room temperature (20–23 °C). The NIR spectrometer was calibrated with benzene ($C_6H_6$) and cyclohexane (c-$C_6H_{12}$) at least twice per day to minimize the influence of variable laboratory conditions [20]. The standard hydrocarbon spectra were used to check the constancy/stability of the spectrometer response during the measurement time (weeks). The spectral range between 9000 and 4500 cm$^{-1}$ (1110–2500 nm) was scanned with a resolution of 8 cm$^{-1}$ (Table 2). Sixty-four (64) transmittance scans were averaged for each sample spectrum. A background spectrum (32 scans) was measured every 45 min. A photometric accuracy of ~0.07% was obtained [28,29]. A cylindrical glass cell (8 mm in diameter) was used throughout the study. Approximately 1 mL of sample was needed for each NIR measurement. NIR spectrum collection was repeated 5 times with cell rotation inside the spectrometer to minimize interferences from the cell or glass defects. The measurement of each sample took less than 3 min. The averaged and background-corrected spectrum was used for subsequent data preprocessing [37–39]. See Refs. [34,37–39] for details.

### 2.2.2. Medium infrared (MIR) spectroscopy

The IR spectroscopy method was used to determine the melanine content in milk products. An analysis was performed using a Tensor 27 FT-IR Spectrometer (Bruker, Germany). A MIRacle ATR add-on device (Bruker, Germany) with a ZnSe crystal (1-reflexion; 1.8 mm) was used for the analysis of the dairy products [33,35]. A spectral range between 500 and 4000 cm$^{-1}$ was scanned with a spectral resolution of 2 cm$^{-1}$ (Table 2). Thirty-two (32) scans were averaged for each spectrum. The measurement was repeated at least 5 times for each liquid and solid sample. See Refs. [33] and Table 2 for details.

### 2.2.3. NIR/MIR spectra preprocessing

Data preprocessing was done according to Ref. [37]. Briefly, prior to building the calibration model, various widely used preprocessing techniques were applied to the data. Nine (9) data pretreatment methods were tested:

- **MC**: Mean Centering.
- **RS** (for ANN only): Range scaling in different intervals: [0.0;1.0], [0.1;0.9], [0.2;0.8], and [0.3;0.7].
- **MSC-MC**: Mean scattering correction followed by mean centering.
- **SNV-MC**: Standard normal variate scaling plus mean centering.
- **SGD1-MC**: First-order Savitzky–Golay derivative followed by mean centering.
- **SGD2-MC**: Second-order Savitzky–Golay derivative followed by mean centering.
- **MC-OSC**: Mean centering followed by the orthogonal signal correction method.
- **SGD1-MC-OSC**: SGD1-MC followed by the orthogonal signal correction method.
- **SGD2-MC-OSC**: SGD2-MC followed by the orthogonal signal correction method.

A detailed discussion of different pretreatment methods for vibrational spectra can be found in Ref. [37]. The SGD1-MC-OSC

**Table 2**
Spectroscopic parameters for each data set and spectral range: near-infrared (NIR) and mid-infrared (MIR).

| | | Unit | Infant formula | | Milk powder | | Milk | |
|---|---|---|---|---|---|---|---|---|
| | | | MIR | NIR | MIR | NIR | MIR | NIR |
| Spectral range | | $cm^{-1}$<br>nm | 500–4000<br>20,000–2500 | 4500–9000<br>2500–1110 | 500–4000<br>20,000–2500 | 4500–9000<br>2500–1110 | 450–3850<br>22,222–2597 | 4500–9000<br>2500–1110 |
| Resolution | | $cm^{-1}$ | 2 | 8 | 2 | 8 | 2 | 8 |
| Optical path | | cm | ATR | 0.8 | ATR | 0.8 | ATR | 0.8 |
| Number of scans | Spectrum | – | 32 | 64 | 32 | 64 | 64 | 72 |
| | Background | – | 32 | 64 | 32 | 64 | 64 | 72 |
| Background measurement | | – | Before each sample | Before and after each sample | Before each sample | Before and after each sample | Before each sample | Before and after each sample |
| Number of spectra per sample | | – | 5 | 5–7 | 5 | 5 | 5 | 5–7 |

ATR (ATR-MIR) = Attenuated Total Reflectance Mid-Infrared Spectroscopy.

method produced the best results (see below) and was used for data pretreatment, unless otherwise specified.

*2.2.4. Calibration models (short introduction)*

*PLS, OPLS, Poly-PLS, and ANN*: See Refs. [38,39] for detailed discussions of different multivariate methods and algorithms: partial least squares regression/projection to latent structures, orthogonal projection to latent structures (OPLS/O-PLS), polynomial partial least squares regression (Poly-PLS), and artificial neural networks. Spline-PLS method has not been applied due to its relatively high computational cost (long optimization procedure) and minimally superior accuracy over the simpler Poly-PLS method [38]. The same can be said about principal component regression (PCR) in comparison with PLS method [38].

*LS-SVM and SVR*: Support vector machines (SVMs) were initially developed by Vapnik [50,51] as a binary classification tool. SVMs are based on some "beautifully simple ideas" [52] and provide a clear intuition of what learning from examples is all about. Intuitively, an SVM model is a representation of the calibration sample set as vectors in space mapped so that the samples from the separate categories are divided by a clear gap that is as wide as possible. New samples from cross-validation or a test set are then mapped into that same space. Based on which side of the gap between classes they fall, they are predicted to belong to one category or another. SVMs show high performance in practical applications when solving sophisticated classification problems [50–52].

The principles of SVM can easily be extended to regression tasks. For a detailed in-depth theoretical background on SVMs for both classification and regression, see Refs. [50–52]. No math will be used in the following text; see Refs. [53–55] for all necessary equations and formalism.

Similar to the approach of ordinary least squares (OLS) and PLS, support vector regression (SVR) also finds a linear relationship between the regressors (input variables, $X$) and the dependent variables ($y$) [53]. The cost function (the function that is minimized to obtain the best regression model) consists of a two-norm penalty on the regression coefficients, an error term multiplied by the error weight, $C$, and a set of constraints. Using this cost function, the goal is to simultaneously minimize both the coefficients' size and the prediction errors (the function's smoothness and accuracy). The first point is important because large coefficients might hamper generalization due to their tendency to cause excessive variance [53].

In SVR, the prediction errors are penalized linearly with the exception of a deviation below a certain value, $\varepsilon$, according to Vapnik's $\varepsilon$-insensitive loss function. Only predictions deviating more than $\varepsilon$ ($|y - y_{pred}| > \varepsilon$, where $y_{pred}$ is the SVR model prediction) are taken into account. The objects with prediction errors larger than $\varepsilon$ are called "support vectors", and only these vectors determine the final prediction of the SVR model. Because only the inner product is used in all calculations, it is possible to use kernel functions, or kernels, that enable nonlinear regression in a very efficient way. The values of $\varepsilon$ and the parameter $C$ have to be defined by the user; both are problem- and data-dependent [51,53].

The ideology of the LS-SVM method is very close to that of SVR, but in this case, the more usual sum of the squares of the errors is minimized, and no $\varepsilon$-based selection is made between the samples. This is a general feature of least squares (LS) methods [55]. This can make the final model more accurate and less computationally expensive; see Ref. [55] for extra details. Parameter $\gamma$, the analog of parameter $C$ in the SVR model, controls the smoothness of the fit.

As described above, SVM-based regression techniques solve many of the intrinsic ANN problems, such as its stochastic nature, the necessity to repeat network training many times, and the nonuniqueness of the final ANN solution. Thus, SVR and LS-SVM are interesting and promising alternatives to ANN. Note that the most important advantage, namely, the possibility of building a nonlinear model, is still valid with the SVM regression. Here, we will try to understand the extent to which SVM-based techniques can substitute ANN-based approached techniques in real-world (industrial) applications.

*2.2.5. Data analysis. Error model optimization.*

*Software and computing*: The initial spectra were digitized using the special software complex created by one of the authors (B.R.M.) [38]. After digitization, each spectrum was represented as a $1 \times 563$ ($1 \times 1750$) vector for NIR (MIR) spectral data: the length of the vector was defined by the spectrometer resolution. The software package MATLAB 2010a (Mathworks Inc., Natick, MA) along with the Statistics Toolbox and the Neural Network Toolbox were extensively used in designing and executing multivariate procedures. Standard programs were modified and extended by B.R.M. Self-written software was used for spectral preprocessing [38,39].

*Outlier detection*: All results are reported for outlier free data sets. Outliers were detected by principal component analysis (PCA) using standard schemes [50].

*Model efficiency (calibration error) estimation*: Six hundred and twenty-one (621), five hundred and ninety-four (594), and five hundred and ninety-four (594) samples were used for calibration model optimization and training for infant formula, milk power, and liquid milk analyses, respectively. The root mean squared error of cross-validation (RMSECV) was used to characterize the prediction capacity of the created model and to optimize its parameters [38,39]. The nine-fold cross-validation (CV) procedure was applied for model optimization [37–39]. The root mean squared error of prediction [37] was used to check the multivariate model accu-

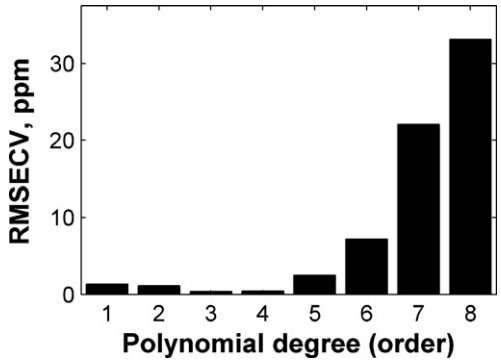


**Fig. 1.** The optimization of the polynomial PLS (Poly-PLS) method by varying the polynomial degree ($n$ = 1–8). The lowest RMSE of cross-validation (CV) is reached with $n$ = 3.

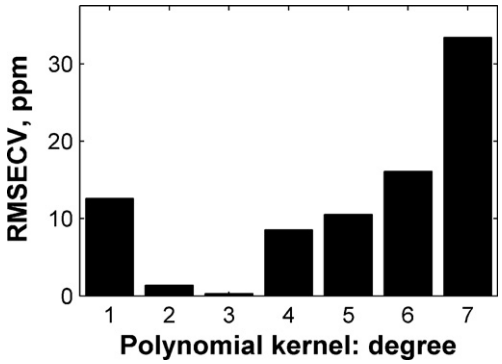


**Fig. 2.** The optimization of the least squares support vector machine (LS-SVM) method by varying the degree of the polynomial kernel ($n$ = 1–7). The lowest RMSE of cross-validation is reached with $n$ = 3.

racy on an independent data set (69, 66, and 66 samples of infant formula, milk power, and liquid milk, respectively).

The test set samples were randomly extracted from the total samples set. Afterwards their representativity (melamine concentration limit, diary product parameters, etc.) was checked to exclude the "extrapolation task" for the final models. The same procedure was used in Refs. [20,22,23,32,33].

*Model optimization*: To truly compare the different multivariate regression methods, the efficiency of the best possible model should be found (see also below). Because of the dependence of the calibration model efficiency on its parameters, the following parameters were varied (optimized) [38]:

- **PLS** (or **Linear-PLS**): Number of latent variables (LV = 1–20).
- **OPLS/O-PLS**: Number of latent variables (LV = 1–20).
- **Poly-PLS**: Number of latent variables (LV = 1–20) and degree of the polynomial ($n$ = 1–10), see Fig. 1.
- **ANN**: Number of latent variables/input neurons (LV/IN = 1–18), number of hidden neurons (HN = 1–18), and transfer function of the hidden layer ($f(\mathbf{x})$ = {*logsig*; *tansig*}). ANN training parameters are shown in Table 1 of Ref. [38].
- **LS-SVM**: The regularization parameter ($\gamma$), determining the trade-off between the fitting error minimization and the smoothness of the estimated function, and the kernel-related parameters (e.g., $\sigma$ or $\sigma^2$ for the RBF kernel). See above for the parameter definitions and other clarifications.
- **SVR**: The error weight ($C$), maximal error value ($\varepsilon$), and kernel-related parameters. The same set of kernels (linear, polynomial, and radial basis function (RBF)) was used for SVR and LS-SVM model building. See Table 4 in Ref. [24] for a detailed list of parameters. See above for the parameter definitions and other clarifications.

Polynomial kernels (default) were found to produce the lowest cross-validation errors in all cases studied (Fig. 2).

## 3. Results and discussion

### 3.1. Method optimization

Figs. 1 and 2 show the procedure used to optimize the Poly-PLS and LS-SVM multivariate methods, respectively [50]. Cross-validation error was minimized in each case. It is interesting to note that the same degree of a polynomial (3) of the final models was found for both methods. The quality of the Poly-PLS method was found to be less dependent on the choice of $n$ (compare Figs. 1 and 2). Other chemometric algorithms (e.g., PLS, OPLS, ANN,) were optimized in a similar manner [38].

### 3.2. The separation into two calibration ranges: 'low' vs. 'high'

It is difficult to make a general calibration model for a wide range of melamine content in dairy products, as we have in our sample (0–2000 ppm). If the whole data set is used for calibration model building, the final model will be biased toward the accurate prediction of samples with high melamine content, due to the minimization of mean squared error (MSE) in most multivariate algorithms ($1000^2 \gg 100^2$) [50]. Thus, the consequence will be low model accuracy for diary samples with low melamine content and, thus, a high limit of detection. Such a result is unacceptable from a perspective of food product quality control [1].

To solve this problem the samples were separated into two subranges, named 'low' and 'high', with melamine content below 17.3 ppm (14.6 ppm) and between 17.3 and 2000 ppm (14.6–2000 ppm) for infant formula (milk powder/liquid milk). Separate calibration models were built for each range. See Table 1 for details. The accuracy of both the 'low' and 'high' multivariate models was checked to ensure that the border values (17.3/14.6 ppm) were not worse than other regions.

### 3.3. Multivariate methods comparison

Fig. 3 represents the results of the multivariate methods comparison for dairy products with low melamine content. For both spectroscopic ranges and all milk products, the same trend is well seen: linear calibration methods (PLS and OPLS) show a much larger prediction error, exceeding 1 ppm. The average error of the PLS/OPLS methods is $1.31 \pm 0.07$ ppm ($\sigma$), while the error of the Poly-PLS, ANN, and SVM-based methods is almost 5 times smaller ($0.28 \pm 0.05$).

Thus, one can state that there is definitely a nonlinear dependence between the infrared spectrum of the milk samples and melamine content, even in a low concentration range [1,36–39]. This nonlinearity can be corrected by a rather simple Poly-PLS algorithm [37].

Almost identical results are observed for all dairy products in both frequency ranges, with the lowest prediction errors of $0.25 \pm 0.04$ ppm for the SVM-based methods. The results for the ANN method of almost the same quality can be obtained. The SVR method is slightly superior to the LS-SVM method, but the difference is negligible—below the reference method/preparation accuracy ($0.01 \pm 0.01$ ppm). Thus, both methods are equally effective and lead to LOD of $0.76 \pm 0.11$ ppm [1].

Fig. 4 shows the same data as Fig. 3 but for the higher melamine concentration range. It is clear from Fig. 4 that linear models are still not accurate. However, a different trend is observed with high concentrations: the Poly-PLS model, being accurate for low-content

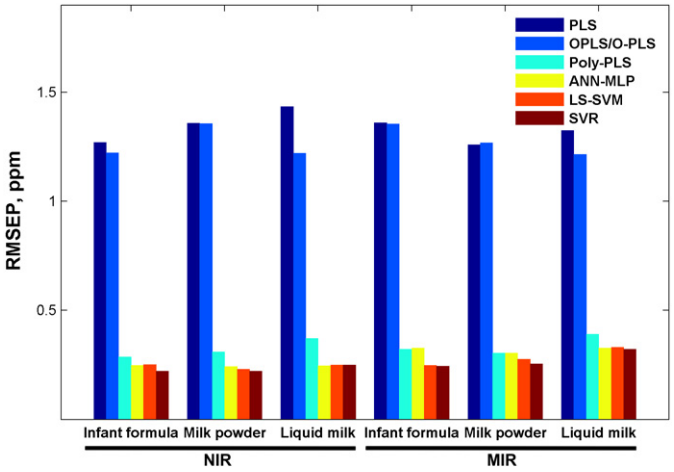


**Fig. 3.** The comparison of the performance of linear and nonlinear calibration techniques—namely, partial least squares regression (PLS), polynomial PLS version (Poly-PLS), artificial neural network (ANN), support vector regression (SVR), and least squares support vector machine (LS-SVM)–for the prediction of melamine content in dairy products from mid-infrared (MIR) and near-infrared (NIR) spectra. The model was created for three milk products: infant formula, milk powder, and liquid milk. The root mean squared errors of predictions (RMSEP) are reported. The samples with low melamine concentration were analyzed (see text for the details).

region (Fig. 3), is not accurate for higher concentrations of the target substance. The ratio between the RMSEP of Poly-PLS and the lowest RMSEP has increased from $1.3 \pm 0.1$ to $2.3 \pm 0.6$.

Thus, for higher melamine concentration, the polynomial modification of the PLS method is not able to fully account for spectral nonlinearity. This effect can be due to the higher influence of intermolecular interactions between melamine and matrix molecules, which are able to slightly shift vibrational frequencies and modify spectral intensities [28–30,56–58].

The ANN, SVR, and LS-SVM methods are able to solve this problem. All of them are able to produce accurate results with an RMSEP of $6.1 \pm 0.9$ ppm.

Note that for high melamine content, NIR spectroscopy is superior to MIR spectroscopy [22,23,36,59].

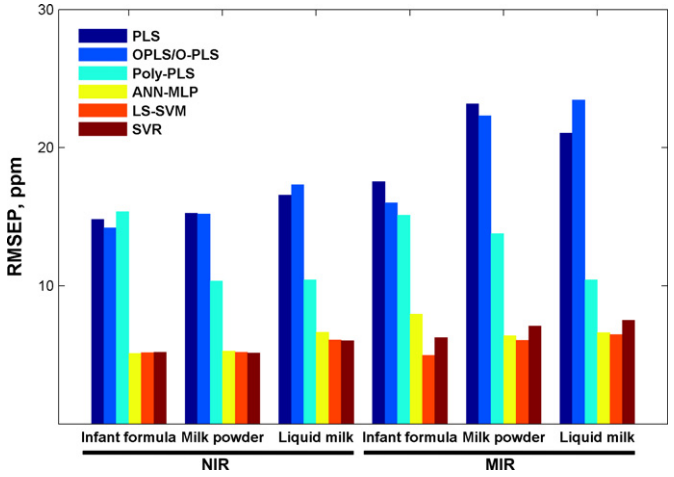


**Fig. 4.** The comparison of the performance of linear and nonlinear calibration techniques—namely, partial least squares regression, polynomial PLS version (Poly-PLS), artificial neural network, support vector regression, and least squares support vector machine (LS-SVM)—for the prediction of melamine content in dairy products from mid-infrared and near-infrared spectra. The model was created for three milk products: infant formula, milk powder, and liquid milk. The root mean squared errors of predictions are reported. The samples with high melamine concentration were analyzed (see text for the details).

### 3.4. Comparison with available literature data

Currently, the FDA uses a liquid chromatography-triple-quadrupole tandem mass spectrometry (LC-MS/MS) method to detect residues of melamine in dry infant formula [60]. Although this method provides limits of detection as low as 0.25 ppm, the sample preparation and cleanup procedures are time consuming and laborintensive.

The spectroscopic method for melamine detection based on mid- and near-infrared spectroscopy provides a much quicker and low-cost procedure with a detection limit of only $0.76 \pm 0.11$ ppm. Even though this value is larger than the LC value by a factor of three, it is still below 1 ppm [1]. That means that the method can be applied in practice to detect the adulteration of dairy products by melamine [1].

Figs. 3 and 4 show that the influence of matrix is rather limited, and different milk products can be analyzed with equal efficiency. Thus, NIR spectroscopy is robust enough for real-world applications. Additional research is needed to clarify the possibility of melamine detection by infrared spectroscopy in other food products.

The presented results (Figs. 3 and 4) can be compared with NIR spectroscopy of petroleum products (e.g., gasoline), as reported by Balabin et al. in 2007 [38]. Note that the hierarchy of the multivariate methods was different in that case: the Poly-PLS method was less effective than the ANN one. Unfortunately, no information on SVM-based multivariate methods was provided [38,39]. The smaller spectral nonlinearity seems to have been observed in a hydrocarbon mixture.

Unfortunately, up-to-date, not many papers compare several chemometric algorithms to find the most appropriate one for this or that task. Thus, it is difficult to make a general comparison of dairy products with other chemical systems to understand the reasons for good or bad (accurate or inaccurate) behavior. We can just state that the ANN and SVM-based models were found to be the most accurate for the melamine calibration task.

However, further work is needed to enlarge the sample banks and the spectral libraries to get a real insight into the chemical reasons for the success or failure of the NIR/MIR + MDA strategy.

### 3.5. Method comparison. General remarks

Here, we confirm the results of Ref. [36] for milk powder and extend them to other dairy products. Both infrared-based methods (MIR and NIR) can be successfully applied to detect melamine in dairy products (infant formula, milk powder, and liquid milk) with a limit of detection below 1 ppm. The simplicity of the sample preparation procedure and recent progress in hand-held near-infrared devices makes the proposed technique promising for real-world applications.

At least preliminary food quality control can be done by IR methods with subsequent confirmation of sample contamination by chromatographic methods. This will greatly decrease the number of samples that are needed for an analysis of the quality and safety of food. The same technique is applied today in the petroleum industry for the online analysis of petroleum refining products and petrochemicals [20,22,23].

The possibility of online (real-time) quality control by NIR spectroscopy is also interesting for the practical implementation of the proposed method. A correct signal preprocessing procedure makes the sample collection unnecessary [37].

## 4. Conclusions

The following conclusions can be drawn:

(1) Infrared spectroscopy (MIR or NIR) is an effective tool to detect melamine in dairy products, such as infant formula, milk powder, or liquid milk.
(2) The limit of detection below 1 ppm can be reached if a correct multivariate algorithm is used for spectrum analysis (LOD = 0.76 ± 0.11 ppm).
(3) The relationship between the MIR/NIR spectrum of milk product and melamine content is nonlinear. Thus, nonlinear regression methods, such as Poly-PLS, ANN, SVR, or LS-SVM, are needed to correctly predict the 2,4,6-triamino-1,3,5-triazine content.
(4) The Poly-PLS method is only effective for low concentrations of melamine in milk samples (<15 ppm).
(5) Mid- and near-infrared spectroscopy can be regarded as a quick, sensitive, robust, and low-cost method for liquid milk, infant formula, and milk powder analysis.

Additional research is needed to clarify the possibility of rapid melamine detection by vibrational spectroscopy in other food products [59,61].

## Acknowledgement

B.R.M. is grateful to the ITERA International Group of companies for a scholarship. Bruker Optics Inc. (Moscow, Russia) is acknowledged for the use of their equipment in the preliminary studies.

## References

[1] M. Lin, Front. Chem. Eng. China 3 (2009) 427–435.
[2] B. Liu, M. Lin, H. Li, Sens. Instrum. Food Qual. 4 (2010) 13–19.
[3] J.R. Ingelfinger, N. Engl. J. Med. 359 (2008) 2745–2748.
[4] H. Xin, R. Stone, Science 322 (2008) 1310–1311.
[5] E.Y.Y. Chan, S.M. Griffiths, C.W. Chan, Lancet 372 (2008) 1444–1445.
[6] K. Burns, S. Kahler, J. Am. Vet. Med. Assoc. 230 (2007) 1784–1785.
[7] S. Wong, M. Chiu, H.K., J. Paediatr. (New Series) 13 (2008) 230–234.
[8] L.J. Mauer, A.A. Chernyshova, A. Hiatt, A. Deeding, R. Davis, J. Agric. Food Chem. 57 (2009) 3974–3980.
[9] E. Cohn, FDA finds more traces of melamine in formula, (CNN, 2008), http://www.cnn.com/2008/HEALTH/11/27/infant.formula.melamine/index.html.
[10] M. Mendoza, Consumer group says FDA melamine guidelines unsafe (2009), http://abcnews.go.com/Business/wireStory?id=66.
[11] World Health Organization, Melamine contamination event, China, 2008, retrieved October, 2009, from http://www.who.int/foodsafety/fs_management/infosan_events/en/index.html.
[12] Grand jury indicts companies for melamine contamination of pet food, JAVMA 2008, 232, 824.
[13] U.S. Food and Drug Administration, Interim safety and risk assessment of melamine and its analogues in foods for humans, http://www.cfsan.fda.gov/~dms/melamra3.html.
[14] Melamine; MSDS 11295, in Alfa Aesar, Ward Hill, MA; http://www.vwrsp.com/msds/10/AAA/AAA11295-0C.pdf.
[15] R.L.M. Dobson, S. Motlagh, M. Quijano, R.T. Cambron, T.R. Baker, A.M. Pullen, B.T. Regg, A.S. Bigalow-Kern, T. Vennard, A. Fix, R. Reimschuessel, G. Overmann, Y. Shan, G.P. Daston, Toxicol. Sci. 106 (2008) 251–262.
[16] L. He, Y. Liu, M. Lin, J. Awika, D.R. Ledoux, H. Li, A. Mustapha, Sensing Instrum Food Qual. Saf. 2 (2008) 66–71.
[17] D.N. Heller, C.B. Nochetto, Rapid Commun. Mass Spectrom. 22 (2008) 3624–3632.
[18] G. Huang, Z. Ouyang, R.G. Cooks, Chem. Commun. 5 (2009) 556–558.
[19] C.M. Karbiwnyk, W.C. Andersen, S.B. Turnipseed, J.M. Storey, M.R. Madson, K.E. Miller, C.M. Gieseker, R.A. Miller, N.G. Rummel, R. Reimschuessel, Anal. Chim. Acta 637 (2009) 101–111.
[20] R.M. Balabin, R.Z. Safieva, E.I. Lomakina, Anal. Chim. Acta 671 (2010) 27–35.
[21] S.B. Kim, C. Temiyasathit, K. Bensalah, A. Tuncel, J. Cadeddu, W. Kabbani, A.V. Mathker, H. Liu, Expert Sys. Appl. 37 (2010) 3863–3869.
[22] R.M. Balabin, R.Z. Safieva, J. Near Infrared Spectrosc. 15 (2007) 343–349.
[23] R.M. Balabin, R.Z. Safieva, Fuel 87 (2008) 2745–2752.
[24] A. Jiye, J. Trygg, J. Gullberg, A.I. Johansson, P. Jonsson, H. Antti, S.L. Marklund, T. Moritz, Anal. Chem. 77 (2005) 8086–8094.
[25] M.R. Monteiroa, A.R.P. Ambrozin, M.S. Santos, E.F. Boffo, E.R. Pereira-Filho, L.M. Lião, A.G. Ferreira, Talanta 78 (2009) 660–664.
[26] J. Workman Jr., M. Koch, B. Lavine, R. Chrisman, Anal. Chem. 81 (2009) 4623–4643.
[27] P. Baptista, P. Felizardo, J.C. Menezes, M.J.N. Correia, Talanta 77 (2008) 144–151.
[28] R.M. Balabin, J. Phys. Chem. A 113 (2009) 4910–4918.
[29] R.M. Balabin, J. Phys. Chem. A 113 (2009) 1012–1019.
[30] R.M. Balabin, J. Phys. Chem. Lett. 1 (2010) 20–23.
[31] J.M. Hollas, Modern Spectroscopy, 4th ed., Wiley, 2004.
[32] R.M. Balabin, R.Z. Syunyaev, S.A. Karpov, Fuel 86 (2007) 323–327.
[33] R.M. Balabin, R.Z. Syunyaev, S.A. Karpov, Energy Fuels 21 (2007) 2460–2465.
[34] F.C.C. Oliveira, C.R.R. Brandao, H.F. Ramalho, L.A.F. Costa, P.A.Z. Suarez, J.C. Rubim, Anal. Chim. Acta 587 (2007) 194–199.
[35] R.M. Balabin, R.Z. Safieva, Fuel 87 (2008) 1096–1101.
[36] C. Lu, B. Xiang, G. Hao, J. Xu, Z. Wang, C. Chen, J. Near Infrared Spectrosc. 17 (2009) 59–67.
[37] P. Felizardo, P. Baptista, J.C. Menezes, M.J.N. Correia, Anal. Chim. Acta 595 (2007) 107–113.
[38] R.M. Balabin, R.Z. Safieva, E.I. Lomakina, Chemometr. Intell. Lab. 88 (2007) 183–188.
[39] R.M. Balabin, R.Z. Safieva, E.I. Lomakina, Chemometr. Intell. Lab. 93 (2008) 58–62.
[40] H. Yang, P.R. Griffiths, J.D. Tate, Anal. Chim. Acta 489 (2003) 125–136.
[41] Y. Li, C.W. Brown, S.-C. Lo, J. Near Infrared Spectrosc. 7 (1999) 55–62.
[42] S. Sekulic, M.B. Seasholtz, Z. Wang, B.R. Kowalski, Anal. Chem. 65 (1993) 835–845.
[43] R.M. Balabin, Chem. Phys. 352 (2008) 267–275.
[44] R.M. Balabin, J. Chem. Phys. 129 (2008) 164101.
[45] R.M. Balabin, J. Chem. Phys. 131 (2009) 154307.
[46] R.M. Balabin, J. Phys. Chem. A 114 (2010) 3698–3702.
[47] A.G. Császár, W.D. Allen, H.F. Schaefer III, J. Chem. Phys. 108 (1998) 9751–9764.
[48] R.M. Balabin, J. Chem. Phys. 132 (2010) 211103.
[49] G. Venkatasami, J.R. Sowa, Anal. Chim. Acta 665 (2010) 227–230.
[50] C.M. Bishop, Pattern Recognition and Machine Learning, Springer, 2007.
[51] V.N. Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag, New York, 1995.
[52] S.R. Amendolia, G. Cossu, M.L. Ganadu, B. Golosio, G.L. Masala, G.M. Mura, Chemometr. Intell. Lab. Syst. 69 (2003) 13–20.
[53] U. Thissen, M. Pepers, B. Ustun, W.J. Melssen, L.M.C. Buydens, Chemometr. Intell. Lab. Syst. 73 (2004) 169–179.
[54] F. Chauchard, R. Cogdill, S. Roussel, J.M. Roger, V. Bellon-Maurel, Chemometr. Intell. Lab. Syst. 71 (2004) 141–150.
[55] U. Thissen, BulentUstun, W.J. Melssen, L.M.C. Buydens, Anal. Chem. 76 (2004) 3099–3105.
[56] R.M. Balabin, J. Chem. Phys. 132 (2010) 231101.
[57] R.M. Balabin, Phys. Chem. Chem. Phys. 12 (2010) 5980.
[58] R.M. Balabin, J. Phys. Chem. B 114 (2010) 15075.
[59] X. Liu, G. Jia, C. Wu, K. g Wang, X. Wu, J. Near Infrared Spectrosc. 18 (2010) 113–120.
[60] U.S. Food and Drug Administration, Determination of melamine and cyanuric acid residues in infant formula using LC-MS/MS, http://www.cfsan.fda.gov/~frf/lib4421.htm.
[61] K. Ai, Y. Liu, L. Lu, J. Am. Chem. Soc. 131 (2009) 9496–9497.