Available online at www.sciencedirect.com



**RESEARCH PAPER** 

Cite this article as: Chinese J. Anal. Chem., 2020, 48(8): e20089-e20097

# Improving Quantitative Accuracy of Ammonia from **Open-Path Fourier Transform Infrared Spectroscopy by Incorporating Actual Spectra into Synthetic Calibration Set** of Partial Least Squares Regression



SONG Qian-Wei, GUO Ming-Quan, SHAO Li-Min\*

Department of Chemistry, University of Science and Technology of China, Hefei 230026, China

Abstract: Partial least squares (PLS) regression is often employed for quantification of ammonia (NH<sub>3</sub>) from open-path Fourier transform infrared (OP/FT-IR) spectra. In this work, PLS models built with both synthetic and actual calibration sets are evaluated, and it is found that the root mean square error of prediction (RMSEP) of the actual model is 86% lower than that of the synthetic model. Although the quantitative accuracy is mediocre, synthetic calibration set is readily constructed, and still required in the beginning when there are insufficient actual OP/FT-IR spectra. However, the RMSEP of the synthetic model is significantly reduced by incorporating some actual spectra into the synthetic calibration set, *i.e.* by constructing a mixed calibration set. Results show that a mixed calibration set composed of three quarters of synthetic and one quarter of actual spectra reduces the RMSEP by 80%, which performs similarly to the actual calibration set. Synthetic, mixed, and actual calibration sets all have pros and cons, and should be utilized according to respective advantages, which results in a novel strategy for NH<sub>3</sub> quantification. The strategy starts with a synthetic model, later builds mixed models by incorporating actual spectra into the synthetic calibration set, and eventually builds an actual model after a large number of OP/FT-IR spectra are collected. This strategy is more accurate than a synthetic calibration set, and more efficient than an actual calibration set, the preparation of which is time-consuming.

Key Words: Quantitative analysis; Calibration; Partial least squares; Open-path Fourier transform infrared spectra

#### Introduction 1

Ammonia (NH<sub>3</sub>) is always an environmental concern as a greenhouse and corrosive gas. It is important to monitor NH<sub>3</sub> emissions around sites such as animal farms and agricultural operations<sup>[1]</sup>. However, NH<sub>3</sub> in air is difficult to be measured accurately because it may be strongly adsorbed on many surfaces, and any technique with a sampling procedure has to prevent the adsorption on the sampling cell walls. Open-path Fourier transform infrared (OP/FT-IR) spectroscopy is ideal to measure NH<sub>3</sub> because the open-path mode obviates the sampling procedure.

OP/FT-IR spectroscopy is widely applied to atmospheric

monitoring of different molecules in various fields<sup>[1-5]</sup>, but the open-path mode makes it more susceptible to environmental interferences and the strong infrared absorption of H<sub>2</sub>O, CO<sub>2</sub>, CH<sub>4</sub>, *etc*<sup>[6]</sup>, OP/FT-IR spectrometer measures accurate vet complicated spectra, which requires advanced chemometric methods. Partial least squares (PLS) regression is often employed to process OP/FT-IR spectra for calculating the concentration of NH3 due to its powerful information extraction capabilities<sup>[7-10]</sup>. In the construction of a PLS model, a key part is the calibration set that should comprise sufficient OP/FT-IR spectra with known concentrations of NH<sub>3</sub>. Moreover, concentrations of calibration samples should be nearly evenly distributed and cover a large range. Such

Received 23 December 2019; accepted 18 June 2020 \*Corresponding author. Email: lshao@ustc.edu.cn

This work was supported by the Fundamental Research Funds for the Central Universities of China (No. wk2060190040).

Copyright © 2020, Changchun Institute of Applied Chemistry, Chinese Academy of Sciences. Published by Elsevier Limited. All rights reserved. DOI: 10.1016/S1872-2040(20)60036-6

calibration spectra are fairly difficult to acquire in a short time since the amount of  $NH_3$  in air cannot be controlled in the open-path mode<sup>[11]</sup>. As a result, calibration spectra for  $NH_3$  are usually synthesized by adding reference spectra of  $NH_3$  with known concentrations to OP/FT-IR spectra measured in pristine air (without  $NH_3$ )<sup>[8,12]</sup>. In this work, it is found that the synthetic calibration set is practically effective, but less accurate than the calibration set that comprises actual OP/FT-IR spectra with known and suitable concentrations of  $NH_3$ . The root mean square error of prediction (RMSEP) of the PLS model built with actual spectra is 86% lower than that built with synthetic spectra.

It takes considerably long time to accumulate actual OP/FT-IR spectra that are suitable to build a PLS model for NH<sub>3</sub> quantification. Before the actual model is available, NH<sub>3</sub> quantification is only carried out with a synthetic model despite its mediocre performance. Our research shows that by incorporating some actual spectra into the synthetic calibration set, the accuracy of the corresponding model is significantly increased. When one quarter of spectra in the calibration set are actual, RMSEP of the model is reduced by 80%.

This work presents a practical strategy for accurate quantification of  $NH_3$  with PLS regression. The strategy starts with a synthetic PLS model to process OP/FT-IR spectra. If OP/FT-IR spectrum is found to be suitable for calibration during the process, it is incorporated into the synthetic calibration set to build a mixed PLS model; eventually an actual PLS model is built when all synthetic spectra in the calibration set are replaced with actual ones. This strategy is more accurate than a synthetic calibration set, and more efficient than an actual calibration set.

# 2 Theory

#### 2.1 Theoretical background

PLS is employed for NH<sub>3</sub> quantification from OP/FT-IR spectra, and the basic theory is summarized as follows<sup>[13]</sup>. Boldface uppercase letters represent matrices, and boldface italic lowercase letters represent column vectors. PLS transforms original variables  $X_{n \times k}$  measured at k variable points with *n* samples into new uncorrelated variables  $T_{n \times p}$ , i.e. latent variable (LV) by the linear combination as shown in Eq.(1), in which coefficient  $\mathbf{W}_{k \times p}$  is a matrix with k rows. The number of LVs, i.e. p, needs certain methods to determine, which is the first step to build a PLS model, and the specific approach is introduced in the next segment. Finally, a PLS model is built by regressing dependent variables  $y_{n \times 1}$  with n concentrations to LVs, and coefficients  $c_{n \times 1}$  are obtained, as shown in Eq.(2). Through a series of substitutions, quantitative relationship between  $y_{n \times 1}$  and  $x_{n \times k}$  can be expressed by regression coefficient  $\boldsymbol{b}_{k \times 1}$ . Vector  $\boldsymbol{e}$  in all equations represents random errors that cannot be fitted into the model.

$$y = \mathbf{T}\mathbf{c} + \mathbf{e} = \mathbf{X}\mathbf{W}\mathbf{c} + \mathbf{e} = \mathbf{X}\mathbf{b} + \mathbf{e}$$
(2)

(1)

Proper number of LVs is crucial for an effective PLS model. Excessive LVs result in an over-fitting model, whereas insufficient LVs make the model under fitting. In this research, Akaike's Information Criterion (AIC) is employed to determine the proper number of LVs. AIC utilizes data from 10-fold cross validation of the calibration set<sup>[14]</sup>, and is calculated with Eq.(3).

T = VW

$$AIC(p) = nln(PRESS) + 2p$$
 (3)

where, p is the number of LVs and n is the sample size of the calibration set. *PRESS* represents the prediction residual error sum of squares, and is calculated with Eq.(4).

$$PRESS = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$$
(4)

where,  $\hat{y_i}$  and  $y_i$  represent the calculated and the actual concentration of NH<sub>3</sub> in the *i*-th calibration spectrum, respectively. When AIC reaches the minimum, the corresponding number of LVs is determined as the optimal one.

The performance of PLS model is evaluated with RMSEP that is calculated with Eq.(5).

$$\text{RMSEP} = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n}}$$
(5)

where,  $\hat{y}_i$  and  $y_i$  represent the calculated and the actual concentration of NH<sub>3</sub> in the *i*-th validation spectrum; *n* is the sample size of the validation set.

# 3 Experimental

#### 3.1 Measurements

The OP/FT-IR spectrometer manufactured by MDA Corp. (Atlanta, GA) contains a Bomem Michelson 100 interferometer, a telescope, a Sterling engine-cooled mercury cadmium telluride (MCT) detector and a cube-corner array retroreflector. The format of spectrum collected by the OP/FT-IR spectrometer is interferogram with a maximum optical path difference of 1 cm (nominal resolution of 1 cm<sup>-1</sup>). Single-beam spectra are computed from interferograms with a zero-filling factor of 8 and medium Norton-Beer (MNB) apodization. Quantification of NH<sub>3</sub> is carried out on spectral data within 1250–750 cm<sup>-1</sup> to avoid strong interference from H<sub>2</sub>O. Calculated concentrations of NH<sub>3</sub> are path-integrated, and the unit of which is ppm-meters or ppm-m.

#### 3.2 Synthetic calibration set

A synthetic calibration set is constructed in three steps<sup>[15]</sup>. At first, 54 single-beam OP/FT-IR spectra without NH<sub>3</sub> are measured in pristine air at different path lengths, temperatures, and relative humidity to encompass characteristics of outdoor air as many as possible. Each transmission spectrum is rationed against a background transmission spectrum measured at a path length of 1 m and converted to absorption spectrum. Then a high resolution ( $0.125 \text{ cm}^{-1}$ ) reference absorption spectrum of NH<sub>3</sub> with the concentration of 75.37 ppm-m is multiplied by 54 different scaling factors and converted to absorption spectrum with the nominal resolution of 1 cm<sup>-1</sup>. Finally, synthetic spectra are obtained by adding reference spectra of NH<sub>3</sub> to the OP/FT-IR absorption spectra of pristine air, and the synthetic calibration set is obtained. Concentrations of NH<sub>3</sub> in synthetic spectra are products of 75.37 ppm-m and corresponding scaling factors, and range from 0 ppm-m to 1400 ppm-m.

#### 3.3 Actual calibration and validation sets

OP/FT-IR spectra containing NH<sub>3</sub> were measured in June and July 2004, and January, March and June 2005 on a dairy farm and fields in southern Idaho. Concentrations of NH<sub>3</sub> in these spectra are roughly determined by the PLS model built with the synthetic calibration set at first. According to calculated concentrations, 54 actual spectra are selected, in which NH<sub>3</sub> concentrations are nearly evenly distributed and cover the range of 0-1400 ppm-m. These spectra constitute an actual calibration set for PLS regression. Then NH<sub>3</sub> concentrations are further determined more accurately by an iterative procedure<sup>[9]</sup>. The procedure starts with a high-resolution (0.125  $\text{cm}^{-1}$ ) reference spectrum of NH<sub>3</sub>, and the concentration of which is 75.37 ppm-m. The high-resolution spectrum is multiplied by a tentative scaling factor and converted to low-resolution (1 cm<sup>-1</sup>) reference spectrum. Then the low-resolution reference spectrum is subtracted from an actual OP/FT-IR spectrum, and the standard deviation of the difference between 970 cm<sup>-1</sup> and 960 cm<sup>-1</sup> is calculated. The process is iterated with different scaling factors until the standard deviation reaches a minimum, and the actual NH<sub>3</sub> concentration in the OP/FT-IR spectrum is determined as the product of 75.37 ppm-m and final scaling factor.

A validation set of 20 actual OP/FT-IR spectra is prepared in the same way with the concentration range of 0-1400 ppm-m.

### 3.4 Mixed calibration sets

Mixed calibration sets are obtained by substituting aforementioned actual OP/FT-IR spectra for some of the 54 synthetic spectra in a synthetic calibration set. Substitutions are in the lower, the higher, and the entire ranges of NH<sub>3</sub> concentration, *i.e.* 0–700, 701–1140 and 0–1400 ppm-m. In each range, substitution proportions of actual spectra are 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40% and 50% (5% is corresponding to 3 spectra). Therefore, there are 30 types of mixed calibration sets. To simulate the real situation and improve reliability, each type of mixed calibration set is constructed 100 times, in which actual spectra randomly substitute their synthetic counterparts. So for each type of mixed calibration set, 100 PLS models are built and corresponding 100 RMSEP values are used to evaluate its performance.

Spectral pretreatments and multivariate PLS calibration are both carried out in MATLAB 2017a (The MathWorks Inc., Natick, MA) on the Windows 10 operating system.

# 4 Results and discussion

# 4.1 Comparison of PLS models built on synthetic and actual calibration sets

Two PLS models are built with synthetic and actual calibration sets, and the optimal numbers of LVs are 8 and 13, respectively, as shown in Fig.1. Usually, the more LVs a PLS model uses, the more complex the calibration set is. The extra LVs of the actual model are probably due to the nonlinearity present in actual spectra. Nonlinearity is not accounted for in synthetic spectra, since they are obtained by linearly adding reference spectra of NH<sub>3</sub> to OP/FT-IR spectra of pristine air.

Validation spectra are processed with both the optimal synthetic and actual PLS models, and the RMSEPs are 62.94



Fig.1 Variation of AIC with increasing number of LVs in PLS models built with (A) synthetic and (B) actual calibration sets. Arrows indicate the minima of AIC

and 8.43, respectively. With the RMSEP being 86% lower, the actual model is much more accurate than the synthetic one. The better performance of the actual model is probably because it accounts for the aforementioned nonlinearity in actual OP/FT-IR spectra.

Model robustness is also investigated with validation spectra. In the investigation, the number of LVs changes from the optimal one (8 for the synthetic model and 13 for the actual one) to other values, and the fluctuation of RMSEPs indicates the model robustness. As shown in Fig.2, when the number of LVs changes from 8 to 2 for the synthetic model, the RMSEP decreases significantly from 62.94 to 23.25. For the actual model, when the number of LVs changes from 13 to 15, the RMSEP decreases from 8.43 to 8.40, which is insignificantly different. Therefore, the actual model is more robust than the synthetic one, and yields less biased results.

#### 4.2 Mixed calibration sets

Although the actual model performs better than the synthetic one, its calibration set is much more difficult to prepare because it takes a long time to find suitable OP/FT-IR spectra with various NH<sub>3</sub> concentrations. So a reasonable solution seems to be starting with a synthetic model. After the OP/FT-IR spectra are measured and processed, some suitable are picked out to build an actual model that is used thereafter. However, we find that by substituting some actual spectra for synthetic spectra of the same NH<sub>3</sub> concentrations, *i.e.* synthetic and actual spectra jointly constructing a mixed calibration set, the corresponding model performs significantly better than the synthetic model. The mixed PLS model is more accurate than the synthetic one, and more convenient than the actual one. It is a good balance of accuracy and efficiency before the final actual calibration set is available and put into use.

PLS models are built with the mixed calibration sets when the substitution of actual spectra is in the NH<sub>3</sub> concentration ranges of 0–700 ppm-m, 701–1400 ppm-m and 0–1400 ppm-m, respectively. Figures 3, 4 and 5 show RMSEPs in the 3 mixed patterns with different substitution proportions, and the average of 100 RMSEPs for each substitution proportion is listed in Table 1. The figures show that the RMSEPs of mixed models are lower than that of the synthetic PLS model. With more actual spectra in the mixed calibration set, the RMSEPs decrease further and more unanimously. From Table 1, it can be found that even for a very low substitution proportion of 5% (only 3 actual spectra in the mixed calibration sets), the average RMSEPs are decreased by more than 50%.

By comparison of these mixed PLS models, it can be observed that the substitution proportions of actual spectra make a difference to the performance of models. When the number of actual spectra is small in mixed calibration set, the model is unstable although the average RMSEP has dropped, especially when the substitution proportions are below about 20% (Fig.3 (A–C), Fig.4 (A–D), and Fig.5 (A–D)). The relatively high analytical precision and accuracy are achieved when the proportion reaches approximately 25%. In such cases, RMSEPs of the mixed models are 80% lower than the synthetic model (Table 1) on average, close to the RMSEP of the actual model.

The substitution concentrations of actual spectra also impact the performance of the mixed PLS models. When the concentrations of actual spectra in the mixed calibration sets are within 0–700 ppm-m or 701–1400 ppm-m, the RMSEPs tend to be stable quickly after the decline, and are always higher than the RMSEP of the actual model, as shown in



Fig.2 RMSEPs of (A) synthetic and (B) actual PLS models built with the number of LVs ranging from 1 to 20. The arrow in (B) indicates the minimum RMSEP

 Table 1
 Average RMSEP of each type of mixed model when the substitution of actual calibration spectra are in different concentration ranges and proportions

Concentration ranges	Substitution proportions								
(ppm-m)	5%	10%	15%	20%	25%	30%	35%	40%	50%
0-700	31.26	18.96	14.04	12.78	12.05	11.94	11.68	11.54	11.57
701-1400	19.02	14.89	14.11	14.00	13.07	12.78	13.03	12.91	12.91
0-1400	26.05	16.58	14.45	13.80	11.20	10.02	9.31	9.05	8.62



Fig.3 (A) to (I) represent 100 RMSEPs of each type of mixed model when actual calibration spectra are in 0-700 ppm-m with substitution proportions of 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40% and 50% (prediction of full validation spectra). Top and bottom dotted lines represent corresponding RMSEPs of synthetic and actual PLS models, respectively



Fig.4 (A) to (I) represent 100 RMSEPs of each type of mixed model when actual calibration spectra are in 701–1400 ppm-m with substitution proportions of 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40% and 50% (prediction of full validation spectra). Top and bottom dotted lines represent corresponding RMSEPs of synthetic and actual PLS models, respectively



Fig.5 (A) to (I) represent 100 RMSEPs of each type of mixed model when actual calibration spectra are in 0–1400 ppm-m with substitution proportions of 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40% and 50% (prediction of full validation spectra). Top and bottom dotted lines represent corresponding RMSEPs of synthetic and actual PLS models, respectively

Table 1. It is because the concentrations of actual spectra in the two mixed pattern only cover half of the entire concentration range. Mixed models are more accurate for the predictions of validation spectra in the same concentration range than for other validation spectral predictions. When the concentrations of actual spectra in mixed calibration sets cover the entire range of 0-1400 ppm-m, the RMSEPs of the mixed models approach that of the actual model with the increase of substitution proportion. So it is desirable that the concentrations of the actual spectra in the mixed calibration set are consistent with, or wider than the validation spectral concentrations.

A further investigation is carried out. The validation set is divided into a low and a high concentration parts, *i.e.* 0–700 ppm-m and 701–1400 ppm-m. Mixed PLS models are built with the actual spectra in 0–700 ppm-m and 701–1400 ppm-m, respectively (referred to as model #1 and #2), and process the two parts of validation spectra. The corresponding RMSEPs are shown in Figs.6–9. By comparing Fig.6 and Fig.7, it is found that for the validation spectra with NH<sub>3</sub> concentrations within 0–700 ppm-m, model #1 yields more accurate results than model #2. The comparison between Fig.8 and Fig.9 indicates that model #2 is better than model #1 in processing validation spectra with NH<sub>3</sub> concentrations of 701–1400 ppm-m. Thus, the consistency of the concentration range between the actual calibration spectra and the validation

spectra accounts for the good performance of mixed models. Actually, this phenomenon is helpful for building mixed models, because it indicates that no requirements are needed for actual spectra selection. The mixed PLS model only needs to be updated with existing actual spectra that have suitable  $NH_3$  concentrations, and the update would not be urgent once the substitution proportion is above 25%.

The mixed calibration set becomes a very cost-effective option due to its predictive accuracy and easy accessibility before the suitable actual calibration set is available. By using synthetic, mixed, and actual calibration sets in the early, middle, and late stages of spectral processing respectively, the quantification of  $NH_3$  in OP/FT-IR spectra becomes more accurate and efficient.

# 5 Conclusions

Calibration sets composed of synthetic spectra are readily obtained, which makes it a convenient choice to construct PLS models to calculate NH<sub>3</sub> concentrations from OP/FT-IR spectra. In this research, we find that both actual calibration set (composed of OP/FT-IR spectra entirely) and mixed calibration set (combination of the synthetic and the actual OP/FT-IR spectra) perform significantly better than synthetic calibration set. It is practically advantageous to utilize the three calibration sets according to their respective strengths in



Fig.6 (A) to (I) represent 100 RMSEPs of each type of mixed model when actual calibration spectra are in 0-700 ppm-m with substitution proportions of 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40% and 50% (prediction of 0-700 ppm-m validation spectra). Top and bottom dotted lines represent corresponding RMSEPs of synthetic and actual PLS models, respectively



Fig.7 (A) to (I) represent 100 RMSEPs of each type of mixed model when actual calibration spectra are in 701–1400 ppm-m with substitution proportions of 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40% and 50% (prediction of 0–700 ppm-m validation spectra). Top and bottom dotted lines represent corresponding RMSEPs of synthetic and actual PLS models, respectively



Fig.8 (A) to (I) represent 100 RMSEPs of each type of mixed model when actual calibration spectra are in 0-700 ppm-m with substitution proportions of 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40% and 50% (prediction of 701-1400 ppm-m validation spectra). Top and bottom dotted lines represent corresponding RMSEPs of synthetic and actual PLS models, respectively



Fig.9 (A) to (I) represent 100 RMSEPs of each type of mixed model when actual calibration spectra are in 701–1400 ppm-m with substitution proportions of 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40% and 50% (prediction of 701–1400 ppm-m validation spectra). Top and bottom dotted lines represent corresponding RMSEPs of synthetic and actual PLS models, respectively

different stages of processing OP/FT-IR spectra, *i.e.* using synthetic calibration set in the beginning, actual calibration set eventually, and mixed calibration sets in between. This strategy is a good balance of accuracy and efficiency in the quantification of  $NH_3$  from OP/FT-IR spectra.

# References

- Bjorneberg L D, Leytem B A, Westermann T D, Griffiths R P, Shao L, Pollard J M. *Trans. ASABE*, 2009, 52(5): 1749–1756
- [2] Goode J G, Yokelson R J, Susott R A, Ward D E. J. Geophys. Res. Atmos., 1999, 104(D17): 21237–21245
- [3] Newman A R. Anal. Chem., 1997, 69(1): 43A-47A
- [4] Richter D, Erdelyi M, Curl R F, Tittel F K, Oppenheimer C, Duffell H J, Burton M. Opt. Lasers Eng., 2002, 37(2): 171–186
- [5] Smith T E L, Wooster M J, Tattaris M, Griffith D W T. Atmos. Meas. Tech., 2011, 4(1): 97–116
- [6] Marshall T L, Chaffin C T, Hammaker R M, Fateley W G.

Environ. Sci. Technol., 1994, 28(5): 224A–232A

- [7] Hart B K, Berry R J, Griffiths P R. Environ. Sci. Technol., 2000, 34(7): 1346–1351
- [8] Shao L, Griffiths P R, Leytem A B. Anal. Chem., 2010, 82(19): 8027–8033
- [9] Shao L, Liu B, Griffiths P R, Leytem A B. Appl. Spectrosc., 2011, 65(7): 820–824
- [10] Lin C H, Johnston C T, Grant R H, Heber A J. Atmos. Meas. Tech., 2019, 12(6): 3403–3415
- [11] Esler M B, Griffith D W, Wilson S R, Steele L P. Anal. Chem., 2000, 72(1): 206–215
- [12] Jiang J H, Berry R J, Siesler H W, Ozaki Y. Anal. Chem., 2002, 74(14): 3555–3565
- [13] Li H, Liang Y, Xu Q, Cao D. Anal. Chim. Acta, 2009, 648(1): 77-84
- [14] Li B, Morris J, Martin E B. Chemome. Intell. Lab. Syst., 2002, 64(1): 79–89
- [15] Shao L, Griffiths P R. Environ. Sci. Technol., 2007, 41(20): 7054–7059