Chemical Research in Toxicology

Article

Cite This: Chem. Res. Toxicol. 2018, 31, 964–973

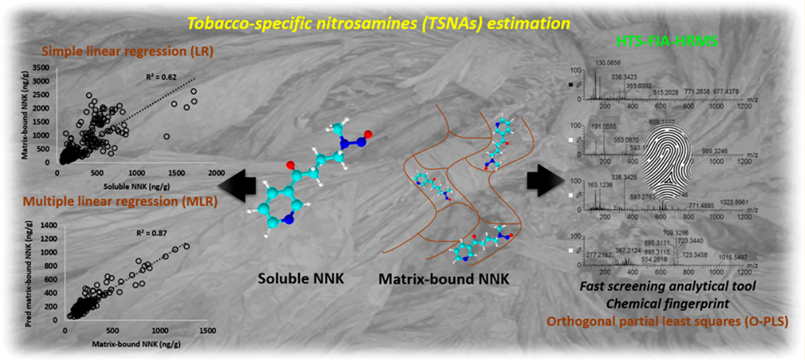
pubs.acs.org/crt

Innovative Approaches for Estimating the Levels of Tobacco-Specific Nitrosamines in Cured Tobacco Samples

Samuel Kaiser, Frederico L. F. Soares, Jorge A. Ardila, Marcelo C. A. Marcelo, Jailson C. Dias, Liliane M. F. Porte, Carlos Gonçalves, Oscar F. S. Pontes, and Guilherme P. Sabin*

British American Tobacco (BAT), Cachoeirinha, Rio Grande do Sul 94930-000, Brazil

Supporting Information



ABSTRACT: Tobacco-specific nitrosamines (TSNAs), mainly the 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone (NNK), are known carcinogens. Part of the NNK found in smoke is provided from matrix-bound NNK, and its determination is extremely relevant. However, the reference extraction procedure of matrix-bound NNK is time-consuming and labor-intensive and has a limited analytical capacity. Three different methodologies were proposed to predict matrix-bound NNK: simple linear regression (LR) with soluble NNK; multiple linear regression (MLR) considering soluble NNK and characteristic parameters of the samples; and orthogonal partial least-squares (O-PLS) regression using high-throughput screening by flow injection analysis coupled to high-resolution mass spectrometry (HTS-FIA-HRMS) data. Simple linear regression showed a high influence of matrix and leaf origin. Although an existing linearity trend has been observed ($R^2 = 0.62$) for the global model, higher correlation values were achieved for matrix and country segregation models. Multiple linear regression predicted matrix-bound NNK with more satisfactory efficiency than simple linear regression models. The coefficients of determination were 0.87 and 0.94 for flue-cured Virginia and air-cured Burley, respectively. However, this method has a limited application, since previous information about the sample is required. The proposed method based on HTS-FIA-HRMS and O-PLS has shown the most suitable performance in the prediction of matrix-bound NNK, with errors comparable to the reference method, and a higher throughput. In addition, this approach allows to determine other soluble nitrosamines, namely *N*′-nitrosoanatabine, *N*′-nitrosoanabasine, and *N*-nitrosonornicotine, with relative percentage errors between 5.25 and 11.98%. Therefore, the third approach is the best method for a large number of cured tobacco for accuracy in determination of TSNAs.

1. INTRODUCTION

Tobacco-specific nitrosamines (TSNAs) are a class of compounds that are widely studied due to its harmful properties.[1] Although no significant amounts of TSNAs are found in green tobacco, their levels are highly increased in tobacco smoke. It is well established that the main formation of TSNAs occurs during the process of curing, storage, and fermentation of harvested tobacco leaves.[2] Nitrosamines are products of nitrosation reaction of nicotine and other endogenous alkaloids with NOx during the post-harvest tobacco and curing process,[3,4] where the most abundant nitrosamines compounds are *N*′-nitrosoanatabine (NAT), *N*′-nitrosoanabasine (NAB), *N*-nitrosonornicotine (NNN), and 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone (NNK).[5] The nitrosamines NNK and NNN are classified as group 1 (carcinogenic to humans) according to the International Agency for Research on Cancer (IARC),[6] whereas NAB presents only limited evidence of risk of esophageal tumors. On the other hand, NAT does not present any evidence of potentially carcinogenic properties.[7]

NNK is a lung-specific selective toxicant formed through nitrosation of nicotine and its degradation products, during the latter stages of tobacco curing and fermentation.[8] Its toxicity effect is associated with carcinogenic potential in rodents, inducing lung tumors independently of its administration route.[9−12] In addition, its metabolite, 4-(methylnitrosoamino)-

Received: June 7, 2018
Published: August 16, 2018

ACS Publications

DOI: 10.1021/acs.chemrestox.8b00154

1-(3-pyridyl)-1-butanol (NNAL), also presents adverse health effects related to pancreatic cancer.[13]

Due to its carcinogenic properties, NNK has been widely studied either in tobacco cured leaves and its smoke.[4,14] Yet, most of the studies found divergence of the amount of NNK found in the tobacco before and after its pyrolysis. Recently, Lang and Vuarnoz developed a method to extract NNK from tobacco cured leaves, in which the results of the total NNK extracted were closely related to the reported value of NNK in most of the cigarette smoke analysis.[15] The results reinforced the idea of two sources of NNK in tobacco leaves: one that is labile and soluble in water and another bounded to tobacco natural polymers. Through the proposed method, the authors found that matrix-bound NNK presented in tobacco leaf corresponds to a large proportion of the mainstream smoke, which suggests that most of the NNK found in the mainstream smoke is derived from matrix-bound NNK, rather than soluble NNK.

Therefore, a method to determine both sources of NNK in tobacco cured leaf, specially matrix-bound NNK, is highly needed in order to monitor the total amount of NNK presented in leaves before the commercial cigarettes manufacture and ensure a less toxicant product. The proposed method to determine matrix-bound NNK in tobacco samples is time-consuming, labor-intensive, and expensive and has a limited analytical capacity, restricting its application in routine laboratory analysis with a high demand for analyses. Another work suggested a theoretical methodology in order to comprehend the relationship of these two NNK sources and its total amount in cigarette smoke.[16] But there is still no work proposing an available more efficient methodology to estimate the matrix-bound amount of NNK in harvested and cured tobacco leaves.

The aim of this work was to develop and compare three methods to predict matrix-bound NNK in tobacco samples without the necessity of laborious extraction procedures. Different methodologies of statistical complexity were studied, starting from a simple linear regression methodology to more complex multivariate methods. The first two approaches were based on statistical models and multilinear regression from reference values of soluble NNK and leaf sample proprieties. The last one was a general method to predict the two sources of NNK and its total composition, based on mass spectrometry analysis and chemometric methods using a simple biphase extraction procedure. In addition, the same chemometric method was also applied to predict other soluble TSNAs in cured tobacco, such as NNN, NAB, and NAT.

2. EXPERIMENTAL PROCEDURES

2.1. Materials and Reagents. Standards for NNK, NAB, NAT, NNN, NNK-d_4 (≥98% purity), Tris(hydroxymethyl)aminomethane (Tris) (≥99.8% purity), ammonium acetate (≥99% purity), sodium hydroxide (P.A.), sodium sulfate (P.A.), chloridric acid (≥99% purity), and HydroMatrix peptide hydrogel were purchased from Sigma-Aldrich (USA). The solvents, methanol, chloroform, acetonitrile, dichloromethane, ethyl acetate, and formic acid were UPLC-MS grade purity and were also purchased from Sigma-Aldrich (USA). Ultrapure water was produced by a Milli-Q apparatus (Millipore, MA, USA). Total NNK extraction solution was prepared by solubilizing 6.0600 g of Tris in 750 mL of ultrapure water. HCl was added dropwise until pH = 7.5 was achieved, and the final volume was set to 1000 mL. Mobile phase A, for total NNK reference method, was prepared by solubilizing 0.7708 g of ammonium acetate into 1000 mL of ultrapure water (10 mmol·L^{-1}). Mobile phases for HTS-FIA-HRMS were prepared by diluting 1 mL of formic acid into 1000 mL ultrapure water (mobile phase A) and into 1000 mL acetonitrile (mobile phase B). For mass calibration and lock mass correction, solutions of sodium formate (0.5 mmol·L^{-1}) and leucine encephalin (1 μg·mL^{-1}) (Waters Reference Solutions, USA) were used, respectively.

2.2. Tobacco Samples. In order to predict tobacco-specific nitrosamines, 582 samples, provided from different sources of British American Tobacco (BAT) during 2016 crop year, were evaluated. The 497 samples of cured tobacco leaves were from Bangladesh (n = 65, 13%), Brazil (n = 130, 26%), Bulgaria (n = 10, 2%), India (n = 78, 16%), Indonesia (n = 29, 6%), Mozambique (n = 30, 6%), Turkey (n = 20, 4%), Uganda (n = 40, 8%), USA (n = 25, 5%), and Zimbabwe (n = 70, 14%). Samples from flue-cured Virginia tobacco stems were also collected, 85 samples from Bangladesh (n = 10, 12%), Brazil (n = 25, 28%), India (n = 5, 6%), Indonesia (n = 15, 18%), USA (n = 15, 18%), and Zimbabwe (n = 15, 18%). Thus, the sample data set encompassed 2 types of sample matrices (leaf and stem), 5 tobacco types (flue-cured Virginia, air-cured Burley, sun-cured Oriental, air-cured *Galpão Comum* and sun-air cured *Galpão Comum*), and 10 countries (Bangladesh, Brazil, Bulgaria, India, Indonesia, Mozambique, Turkey, Uganda, USA, and Zimbabwe). All cured tobacco leaves and stems were stored at −20 °C. One week before the analyses, the samples were ground to powder, with a mean diameter of 0.5 mm, and stored at 22 °C until the performance of the analyses.

2.3. Reference Analyses for Soluble Tobacco-Specific Nitrosamines, Total NNK, and Matrix-Bound NNK. Soluble NNK, NNN, NAT, and NAB were determined using an adapted methodology of ISO/TS 22304:2008.[17] Samples (ca. 1 g) were extracted in ultrasonic bath for 20 min, by adding 2.0 mL of NaOH (0.1 mol·L^{-1}), 2.0 g of HydroMatrix, and 25 mL of dichloromethane. Extracts were filtered and dried in a rotary evaporator under reduced pressure at 65 °C. Samples were reconstituted with 1 mL of ethyl acetate containing internal standard (N-nitroso-di-n-hexylamine (NDHA), 800 ng·mL^{-1}). The solution was filtered (0.45 mm PVDF) and added to 2 mL vials containing 0.3 g of sodium sulfate. Analyses were made on Agilent 6890N (Agilent, USA) gas chromatograph (GC) coupled to a thermal energy analysis (TEA) Antek 7090 detector (Agilent, USA). A megabore J&W DB-1 (30 m × 0.53 mm × 5.0 μm) column (Agilent, USA) was used. The injection temperature was set at 250 °C, in splitless mode, 2.0 μL of injection volume, and helium (purity >99.999%) as carrier gas, at 10 psi. The oven temperature program was: initial temperature of 200 °C, increased by 242 °C with a 7 °C·min^{-1} rate, followed to an increase rate of 120 °C·min^{-1} until 280 °C. The detector temperature was 5 °C, with pyrolysis temperature of 350 °C and O_2 rate of 5 mL·min^{-1}. Concentration values were obtained based on the peak area ratio between NNK and its internal standard.

Total NNK was quantified using LC-MS/MS based on the Lang and Vuarnoz[15] method. The extraction procedure was performed using ca. 750 mg of powdered samples and Tris-HCl buffer solution (30 mL; 50 mmol·L^{-1}; pH 7.5) with added internal standard (NNK-d_4). Samples were extracted in autoclave at 130 °C, 275 kPa for 4 h. Extracted solutions were filtered (0.20 μm PTFE) and injected in LC-MS/MS. Analyses were carried out on an Agilent HPLC 1200 system (Agilent, USA) coupled to a triple-quadrupole mass spectrometer (TQMS) (Applied Biosystem API 4000, Agilent, USA), with electrospray ionization (ESI). A Poroshell 120 HPH-C18 column (2.1 × 50 mm; 4 μm) and a Poroshell HPH-C18 (2.1 × 10 mm; 4 μm) guard column (Agilent, USA) were used. The column temperature was set at 50 °C, the autosampler temperature at 10 °C, and with an injection volume of 6 μL. Separation was performed using a gradient between mobile phases A (ammonium acetate 10 mmol·L^{-1}) and B (acetonitrile). The gradient was: 0 min, 95% A, 5% B; 0.86 min, 95% A, 5% B; 5.50 min, 82% A, 18% B; 5.83 min, 4% A, 96% B; 6.50 min, 4% A, 96% B; 7.33 min, 95% A, 5% B; 12 min, 95% A, 5% B; with a flow rate of 500 μL·min^{-1}. Positive ESI was applied under a desolvatation gas flow of 500 L·h^{-1}, capillary voltage of 4500 V, cone voltage of 41 V, and collision energy between 15 and 17 eV. The quantification values of total NNK were calculated based on the

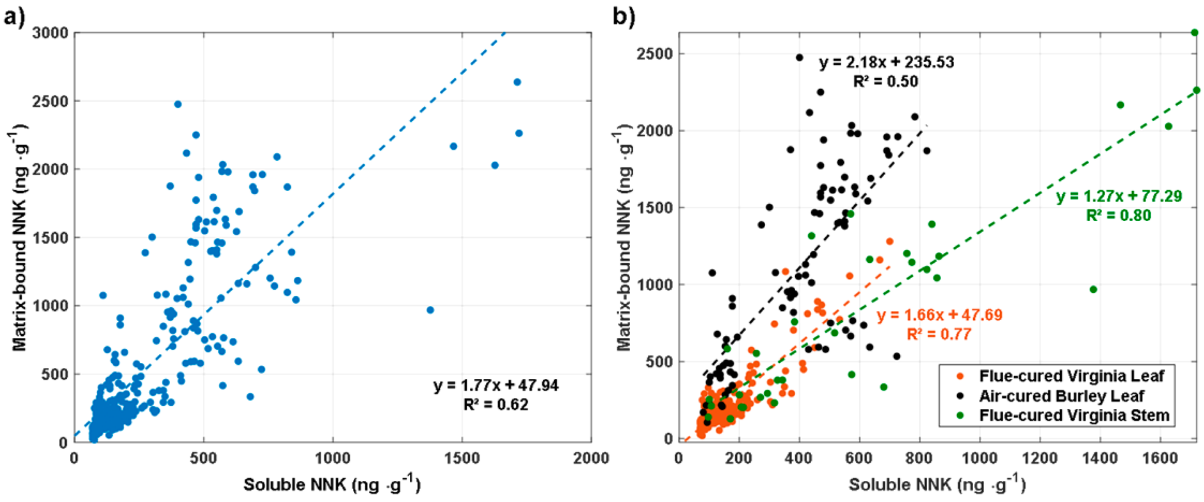


Figure 1. Simple linear regression models for (a) all samples and (b) samples separated by its matrix type.

peak area ratio between NNK and the internal standard. Matrix-bound NNK reference values were obtained from the difference of total NNK and soluble NNK.

2.4. HTS-FIA-HRMS Analysis. Sample preparation and analysis by HTS-FIA-HRMS were performed in accordance with previous work.[18] Briefly, tobacco samples (ca. 200 mg) were extracted using a solution of chloroform:methanol:water (50:25:25 v/v) in ultrasonic bath for 15 min, followed by mechanical shaking for 15 min. The bilayer extracts were achieved after centrifugation at 2500 rpm for 5 min. Two mL of each phase was filtered and diluted 20 times, in acetonitrile for the organic phase and acetonitrile:water (1:1 v/v) for the aqueous phases before the HTS-FIA-HRMS analysis. Each sample was analyzed in both ESI$^+$ and ESI$^-$ modes, for both extraction layers (aqueous and organic), resulting in four fingerprint spectra per sample within 2 min of analysis.

Mass spectra of the samples were obtained in an ACQUITY I-CLASS UPLC (Waters, USA) coupled with SYNAPT G2-Si (Waters, USA) mass spectrometer, where the UPLC system was used only to inject the sample into the mass spectrometer. 300 nL of the sample was injected into the UPLC system with an isocratic mode of 50:50 (v/v) mobile phases A and B for the aqueous phase and 20:80 (v/v) for the organic phase, during 30 s. For ESI$^+$ mode capillary voltage of 3000 V was applied and for ESI$^-$, 2500 V. A source temperature of 120 °C was used as well as a voltage of 80 V, a cone voltage of 40 V, and a desolvatation gas flow of 500 $L \cdot h^{-1}$. Nitrogen was used as a nebulizer, cone, and desolvation gas. Mass spectra were obtained in TOF analyzer using the high-resolution mode (40,000 FW/HM) with a scan time of 0.2 s and a mass range of 50−1000 m/z ratio for the aqueous phase and a range of 50−1200 m/z ratio for the organic phase. To lock mass correction, a leucine encephalin solution (1 $\mu g \cdot mL^{-1}$) was used at an infusion rate of 10 $\mu L \cdot min^{-1}$. Before each batch analysis, the system calibration and detector checking were performed with a sodium formate solution (0.5 $mmol \cdot L^{-1}$) and a leucine encephalin solution (1 $\mu g \cdot mL^{-1}$) (Waters Reference Solutions, USA), respectively.

At the end of each batch analysis, one unique sample of each of three tobacco varieties (air-cured *Galpão Comum*, flue-cured Virginia, and air-cured Burley) were analyzed by HTS-FIA-HRMS and used as control samples. This procedure resulted in 27 control samples for each variety, to assess the model and equipment repeatability.

2.5. Data Analyses. Before any model analysis, the samples with analyte concentration below the limit of quantification of the reference method (70 $ng \cdot g^{-1}$ for soluble NNK LOQ; 34.4 $ng \cdot g^{-1}$ for total NNK LOQ; 60 $ng \cdot g^{-1}$ for soluble NAB LOQ; 80 $ng \cdot g^{-1}$ for soluble NAT LOQ and 80 $ng \cdot g^{-1}$ for soluble NNN LOQ) were removed from the data set. During this exclusion, 250 samples, including all Oriental samples, were removed. After that, a sample selection was performed in the data set based on standardized residuals of the respective response matrix. Samples with a standardized residual (y) lower than or equal to −2.5 and higher than or equal to 2.5 were not considered by the model. 33 samples were considered as outliers in this process, which included all air-cured *Galpão Comum* samples and flue-cured Virginia stem samples from India and Zimbabwe. Therefore, in this work, only flue-cured Virginia and air-cured Burley tobacco leaves and flue-cured Virginia stems were included in the models.

A total of 299 samples resulting from the sample selection were applied to create prediction models for matrix-bound NNK reference values, using two distinct prediction strategies. The first strategy was carried out by using simple linear regression (LR) of reference values for soluble NNK and matrix-bound NNK, segregating models based on matrix, tobacco type, and origin to improve the model predictability. The second approach was based on multiple linear regression (MLR), where sample properties and characteristics were also taken into account to predict matrix-bound NNK. Four qualitative factors were used to characterize the sample: tobacco color (orange or mahogany for flue-cured Virginia, and light brown or brown for air-cured Burley), plant ripeness (immature, mature, and overmature), quality (low, medium, and high), and nicotine content (level 1, 1.45−1.95%; level 2, 1.95−2.45%; level 3, 2.45−2.95%; level 4, 2.95−3.45%; level 5, 3.45−3.95% (w/w) for flue-cured Virginia; and level 2, 1.35−2.05%; level 3, 2.05−2.75%; level 4, 2.75−3.45%; level 5, 3.45−4.15% (w/w) for air-cured Burley). These parameters were in accordance with BAT guidelines for tobacco grading. Soluble NNK amount was inserted into the model as quantitative factor. MLR data analysis was performed using MODDE (version 9.1.1.0, Umetrics, Sweden) software.

To preprocess the HTS-FIA-HRMS data in order to quantify nitrosamines, the same method described previously[18] was applied. Concisely, data obtained from mass spectrometer were imported to MATLAB (MATLAB 2017a, MathWorks, USA), and aligned based on a reference vector, where each fiagram is combined to one mass spectrum. Background and noise were eliminated from the spectra, followed by a normalization step based on the sum of intensities. In the end, each mass spectra (both phases and ion generation modes) were concatenated to generate one single fingerprint for each sample.

The same selected samples were used to create the simple linear regression models, and the multiple linear regression models were used to construct the chemometric model. Furthermore, sample Hotelling (T^2)[19] was used, as well as the distance from the model in X-space (DModX),[20] to detect outliers. Samples that presented a T^2 higher than 95% of confidence level, or a DModX higher than 2.5, were also excluded from the model. At the end of sample selection, 274 samples with 12,367 variables were used to create the models. Data set was analyzed in a MATLAB environment using PLS_Toolbox (version 7.8.1, Eigenvector, USA), where Pareto scaling was applied in mass spectra matrix, followed by orthogonal partial least-squares (O-PLS) regression to develop quantification models. Variable selection was performed based on variable importance in projection (VIP) with a threshold of 1 and a correlation value between data matrix ($\mathbf{X}$) and predictive scores ($\boldsymbol{t}$) from O-PLS (Corr($\mathbf{X}$,$\boldsymbol{t}$)), with an excluding region of correlation values between ±0.4. A Venetian Blind with 7 windowed samples was used to perform model cross-validation in order to determine the number of latent variables and evaluate the model performance. The validation step was evaluated through the bootstrapping methodology. The bootstrapping procedure is based on the random division of samples into calibration and validation sets, both in $\mathbf{X}$ and $\mathbf{y}$ matrices. With these sets, PLS (or O-PLS) models and errors are calculated. This procedure is repeated n times with sample replacement, to ensure that no subsample selection occurred, guaranteeing a real validation error with model robustness.[21−23] 33% of the samples were selected as a validation set, and the procedure was repeated 100 times.

3. RESULTS AND DISCUSSION

3.1. Linear Regression for Estimate Matrix-Bound NNK from Soluble NNK. A linear regression was attempted to predict matrix-bound NNK based on the soluble NNK (Figure 1). It was possible to observe a coefficient of determination (R^2 = 0.62) based on the NNK source concentration (Figure 1-a). A large number of samples with high soluble NNK amount also showed high values for matrix-bound NNK. However, higher correlation values were observed when samples were discriminated according to tobacco types (Figure 1b). Flue-cured Virginia stems and flue-cured Virginia leaves increased their coefficient of determination compared to total samples correlation model, with R^2 of 0.80 and 0.77, respectively. The linear regression models indicated that matrix-bound NNK is strictly related to the tobacco variety and the ratio of soluble NNK to Total NNK. The differences verified in the regression line slopes of flue-cured Virginia, air-cured Burley tobacco leaves, and flue-cured Virginia tobacco stem reinforce this hypothesis. In general, air-cured Burley tobacco showed higher matrix-bound NNK when compared to flue-cured Virginia, both leaf and stem samples, but with a poor coefficient of determination (R^2 = 0.50).

The values of matrix-bound NNK for air-cured Burley were, in most of the cases, 2.2 times higher than the soluble NNK. Although for flue-cured Virginia the values of matrix-bound NNK were lower than air-cured Burley, it was still possible to observe more matrix-bound NNK than soluble NNK, in which leaves contained 1.6 more matrix-bound NNK than soluble NNK and stems contained 1.3 times, in most of the samples. The results showed a good agreement with the results demonstrated by Lang and Vuarnoz[15] and corroborates that a large proportion of the NNK in mainstream smoke derives from matrix-bound NNK.

Samples from different countries presented anomalous behaviors in these models, mainly for air-cured Burley leaf samples. Thus, more discriminative models were made based on the correlation of soluble NNK and matrix-bound NNK, with sample segregation depending on its country of origin. Figure 2 presents linear regression models with sample segregation on variety and country of origin.

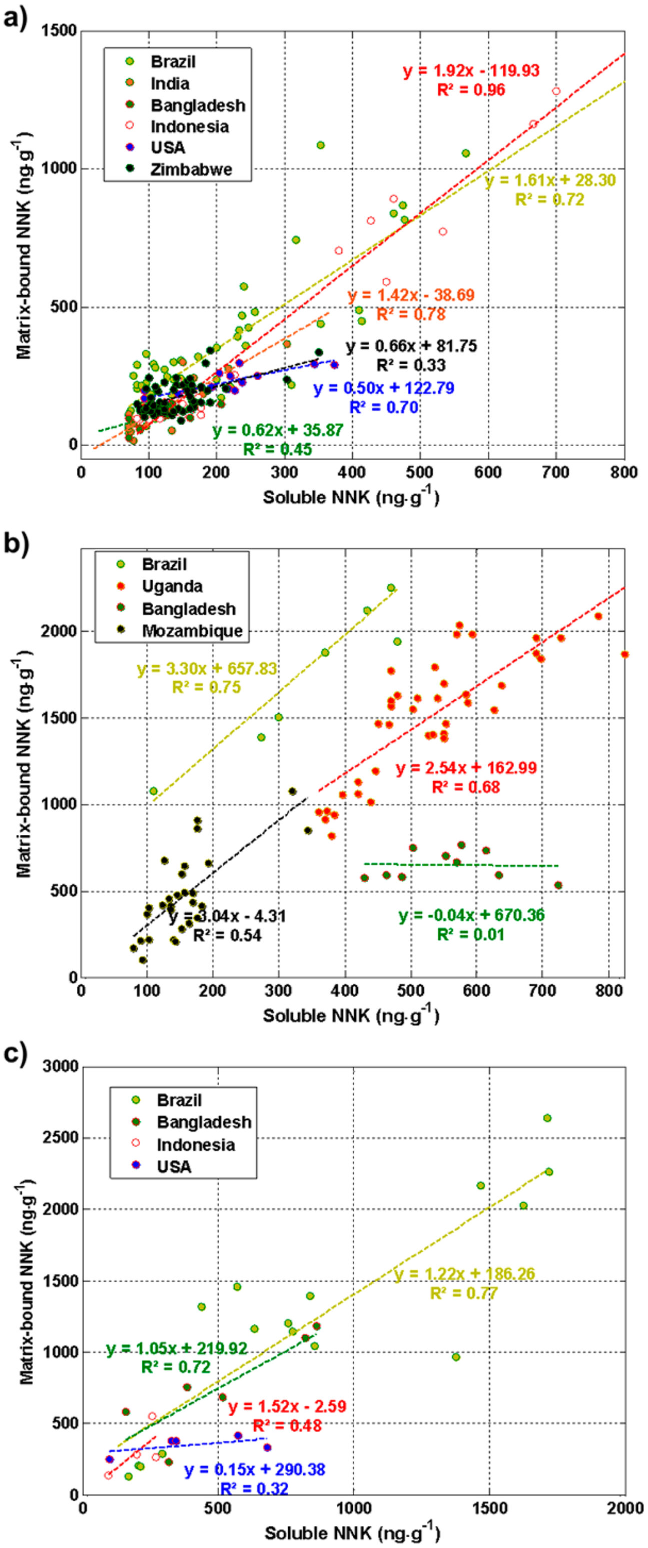


Figure 2. Simple linear regression models based on its country of origin for (a) flue-cured Virginia tobacco leaf, (b) air-cured Burley tobacco leaf, and (c) flue-cured Virginia stem.

Linear regression models with sample segregation on tobacco matrix and origin presented satisfactory coefficients of determination for a simple method, with R^2 higher than 0.65 for most of the models. Samples containing a high amount of matrix-bound NNK, like flue-cured Virginia leaves from India and from Brazil, showed the best coefficients of determination. However, flue-cured Virginia leaves from Bangladesh and Zimbabwe, air-cured Burley leaves from Bangladesh, and flue-cured Virginia stems from the USA and Indonesia presented poor correlation results, with R^2 lower than 0.55. The samples from these countries presented a low amount of total NNK

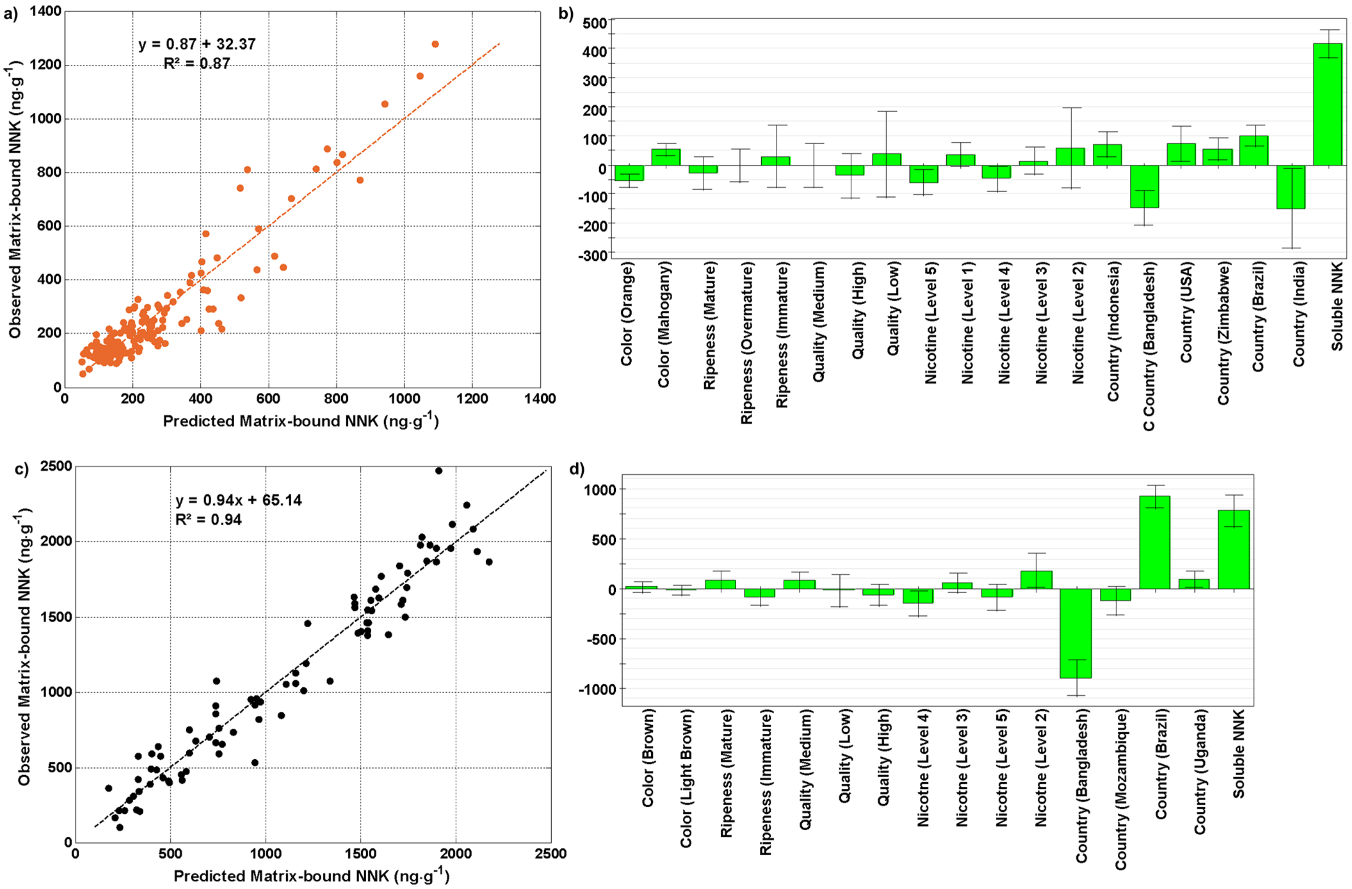


Figure 3. Predictive multiple linear regression models for matrix-bound NNK based on tobacco leaf variety for (a) flue-cured Virginia tobacco and (b) its coefficient parameters and for (c) air-cured Burley tobacco and (d) its coefficient parameters.

and, consequently, a low amount of matrix-bound NNK. Thus, it was difficult to correlate soluble NNK with matrix-bound NNK for these samples. Therefore, in general, the models showed a significant influence of tobacco origin and matrix, in order to correlate soluble and matrix-bound NNK. These results have enabled a better comprehension of the NNK source in different countries, mostly due to its edaphoclimatic conditions and the tobacco varieties used in the studied countries. For example, through the linear regression models, it was possible to observe that the total NNK in the USA samples was relatively lower than the other countries' samples and was more concentrated in the leaves than the stems.

The models of linear regression presented interesting results, but every model was dependent on the sample matrix and origin. Therefore, a linear regression taking into account these parameters in one single model could improve the predictive results. Multiple linear regression (MLR) is a suitable alternative to data modeling, since more than one parameter can be considered simultaneously during the statistical fitting procedures.

3.2. Multiple Linear Regression To Estimate Matrix-Bound NNK from Soluble NNK.

Unlike the linear regression method, MLR evaluated other parameters to build a predictive model with a better performance. Additionally to the sample origin and soluble NNK, other qualitative tobacco grading parameters were also evaluated, such as tobacco color, ripeness, quality, and nicotine content. Due to the reduced number of flue-cured Virginia stem samples, the MLR approach was limited to flue-cured Virginia and air-cured Burley tobacco leaves.

Figure 3 shows the results obtained through MLR for both varieties of tobacco leaves. A better coefficient of determination was reached by means of the MLR methodology, compared to most of the simple linear regression models, with a coefficient of determination for flue-cured Virginia of 0.87 (Figure 3a) and 0.94 for air-cured Burley (Figure 3c). Parameter coefficients and prediction errors permitted to evaluate which factors were significant for the model (Figure 3b,d). Coefficient values with its standard errors and confidence interval are fully detailed in Table S1 for flue-cured Virginia and Table S2 for air-cured Burley. The flue-cured Virginia MLR model demonstrated significant parameters for color, nicotine amount, country of origin, and soluble NNK. The last two parameters were deemed as significant for the linear regression models. The amount of matrix-bound NNK seemed to present a tendency due to its origin. Since different countries provide different tobacco types, these changes can influence the total amount of NNK, either soluble or matrix-bound.

Though NNK is formed through nicotine and its oxidation products via nitrosation process,[2] for flue-cured Virginia samples, the nicotine content was not a factor as significant as soluble NNK, considering its correlation to matrix-bound NNK. Nicotine levels 4 and 5 showed confidence intervals lower than the numeric value of their coefficients. However, the coefficient values were expressively lower than soluble NNK. Even with confidence intervals lower than its numeric

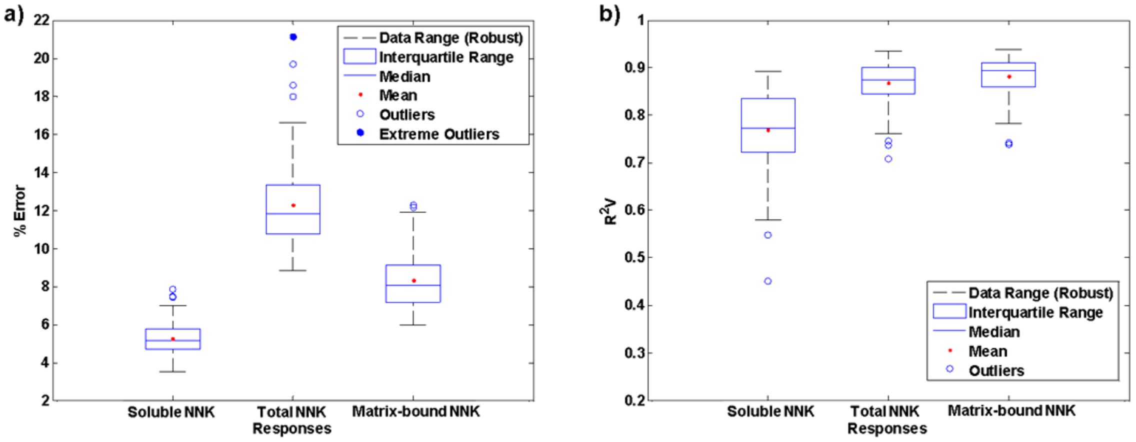


Figure 4. Boxplot of (a) percentage errors based on RMSEV and concentration range and (b) coefficient of determination of validation obtained from 100 validation models resulting from bootstrapping.

coefficient, the effect of nicotine content into the predicted matrix-bound was inversely correlated. Levels 1−2 showed a positive effect, while levels 4−5 showed a negative effect.

It is well-known that stalk position of tobacco leaves is highly correlated with the nicotine content. Probably, the possible explanation for the nicotine effect in the MLR model is the indirect stalk position effect. As matrix-bound NNK is linked in the structural chains of the tobacco plant, the bottom leaves could present higher amounts of structural polymers than the top position of the tobacco plant, and thus the lower position leaves would contain more matrix-bound NNK than soluble NNK.

Color is a parameter related to ripeness and the curing process, and it is well-known that the curing process is an essential parameter altering the total amount of TSNAs.[24] The fact that this parameter is also significant was expected, and it was in accordance with the literature.

The air-cured Burley MLR model demonstrated some influence of nicotine amount (level 2 and 4 were significant factors) and clear significance for country and soluble NNK. Compared to flue-cured Virginia samples, air-cured Burley presented a higher influence of country, as previously indicated by linear regression models. In addition, the color discrimination in air-cured Burley was more complex compared to flue-cured Virginia. Probably, for these reasons, it was not possible to identify color as a meaningful parameter.

Therefore, MLR models showed suitable adjustment to predict matrix-bound NNK from soluble NNK for both flue-cured Virginia and air-cured Burley leaves. For both tobacco types, the coefficients related to tobacco origin were important, as demonstrated in the linear regression models. But for air-cured Burley, its influence was more notable than for flue-cured Virginia. Nicotine amount and color were the other important parameters that could affect the matrix-bound NNK. These observed effects could be confirmed by a more extensive sampling, comprising different tobacco crops and countries of origin.

3.3. Orthogonal Partial Least Squares Regression To Estimate NNK from Chemical Fingerprint.

Multiple and simple linear regression models based on the soluble and matrix-bound amount of NNK presented interesting results, where the MLR methodology could be used to correlate the matrix-bound amount of NNK in tobacco samples through soluble NNK. However, most of these results were regarded as local models that depend on a preview knowledge about sample types (leaf or stem), variety, color, origin, and nicotine content. In addition, some of these parameters are subjective responses that could decrease the model performance depending on the subjective grading process.

The HTS-FIA-HRMS system has already been demonstrated as a useful technique to quantify multi-analytes in several sample matrices with fast analysis.[25−28] Therefore, a methodology to quantify soluble, matrix-bound, and total NNK in cured tobacco samples, regardless of the sample type, origin, or total amount of alkaloids, was proposed by using this method.

O-PLS models were performed with randomly selected 183 samples (66%) for calibration and 91 samples for validation (33%), using VIP as variable selection (with a threshold of 1), followed by a correlation between the $\mathbf{X}$ matrix and its scores ($\mathrm{Corr}(\mathbf{X},\boldsymbol{t})$) from S-plot (with variables excluded region of ±0.4). The root-mean-square errors were calculated for calibration, cross-validation (using Venetian blind with 7 windowed samples), and validation, through random external samples selected during bootstrap. This procedure was done 100 times randomly, by replacing the samples for the validation stage. The mean error values of all 100 models were used to evaluate the methodology.

Figure 4 presents the boxplot of percentage errors (based on mean values for square root mean errors of validation (RMSEV) and the linearity range of the model) (Figure 4a) as well as the coefficient of determination between the reference value and the predicted value during the validation step (R^2V) (Figure 4b) for all 100 models. It is possible to observe that, even replacing the samples for validation, the predicted values remain stable in a close range, with the highest interquartile range of 10.8−13.7 for % error in total NNK and 0.71 to 0.81 for R^2V of soluble NNK. Remaining interquartile values were within 5% of difference for the other monitored results. A few outlier models were observed for all the evaluated responses. As bootstrap randomly removes samples without any selection, in these outlier models, representative samples were probably removed from the calibration data set, which consequently decreased the model performance. Nevertheless, results were still in a satisfactory range of acceptability, and representative samples were not expected to be removed from the data set in the routine analysis. These results demonstrate the model performance and robustness, ensuring that no overfitting was observed in the models.

Table 1 presents the average results obtained from O-PLS bootstrapping models for total NNK, soluble NNK, and matrix-bound NNK, with their respective values of range response (minimal and maximum response values); the coefficient of determination between the reference value and predicted values by calibration model (R^2), predicted by the cross-validation model (R^2CV), and the validation model (R^2V); the root-mean-square error of calibration (RMSEC) of cross-validation (RMSECV) and validation (RMSEV); % error based on the RMSEV and the range of the model; latent variables (LV) and the number of selected variables (V_{sel}). Figure S1 presents the real values against the predicted values for the three NNK models.

All models presented suitable results, with high coefficients of determination, between 0.73 and 0.89 for cross-validation samples and 0.77 to 0.88 for validation samples. Errors of calibration, cross-validation, and validation were close, with results slightly better for calibration. These results reinforce that no overfitting occurred in the models, once external samples presented similar results to the calibration ones. The total NNK was the response with the highest relative error, 11.98%. The remaining ones presented percentage errors below 9%. Soluble NNK presented lower results of R^2V as compared to models for prediction of total and matrix-bound NNK. However, this was expected, as this parameter presented the lowest concentrations, with the value of several samples close to the reference LOQ.

The permutation test was performed in the model without splitting sample data set in calibration and validation tests to make sure that no overfitting was achieved. The **y** matrix with the value response was permuted 1000 times for each response. Models were evaluated by the R^2 and R^2CV of permuted values compared to the original **y** matrix. From histograms of correlation values obtained during the permutation test, no evidence of overfitting was verified, as previously observed through the bootstrapping methodology (Figure S2). In addition, control samples were analyzed and evaluated in a control chart to ensure repeatability (Figure S3). All predicted values presented a response lower than its mean ± 2 times the standard deviation, with a relative standard deviation (RSD) of 9.58% for air-cured *Galpão Comum* (Figure S2a), 11.09% for flue-cured Virginia (Figure S2b), and 11.62% for air-cured Burley (Figure S2c). These results were satisfactory for all the other responses, and better precision can be achieved with more replicates.

Table 1. Summarized Results for Quantitation of Soluble, Total, and Matrix-Bound NNK by HTS-FIA-HRMS and O-PLS Proposed Methods

analyte	soluble NNK	total NNK	matrix-bound NNK
min (ng·g^{-1})[a]	70.00	168.08	71.41
max (ng·g^{-1})[b]	1720.00	4350.75	2637.42
R^2[c]	0.88	0.93	0.93
R^2CV[d]	0.77	0.89	0.89
R^2V[e]	0.77	0.87	0.88
RMSEC (ng·g^{-1})[f]	89.29	211.99	156.89
RMSECV (ng·g^{-1})[g]	123.07	274.40	190.62
RMSEV (ng·g^{-1})[h]	134.79	315.09	213.87
% error[i]	5.25	11.98	8.33
LV[j]	7	7	6
V_{sel}[k]	930	1106	1401

[a]Minimum value of concentration range. [b]Maximum value of concentration range. [c]Coefficient of determination of calibration values. [d]Coefficient of determination of cross-validation values. [e]Coefficient of determination of bootstrapping validation values. [f]Root-mean-square error of calibration. [g]Root-mean-square error of cross validation. [h]Root-mean-square error of bootstrapping validation. [i]Percentage error based on RMSEV and concentration range. [j]Number of latent variables. [k]Number of selected variables.

To comprehend the calibration models and make sure that relevant information was maintained during the variable selection, the covariance and correlation of the score vectors resulting from the last model of O-PLS and the original data[29] were calculated. Figure 5 shows the projection of these two parameters for the first latent variable of O-PLS, also known as the S-plot. Some variables with high importance to the model (that presented a high covariance and high correlation with the analyte concentration) were highlighted in the figure with a different color, and with its monoisotopic *m*/*z* ratio.

Specific flavonoids, like rutin ($[M-H]^-$ 609.1449 *m*/*z* ratio), and caffeoylquinic acid derivatives, like chlorogenic acid ($[M-H]^-$ 353.0869 *m*/*z* ratio), showed a high negative correlation and covariation with the matrix-bound and total NNK level in cured tobacco samples (Figure 5), both of the polyphenols chemical class. On the other hand, the oxidative products derived from nicotine, like pseudooxynicotine (PON) and oxynicotine, both isomers with $[M+H]^+$ 179.1189 *m*/*z* ratio, showed a high positive correlation and covariation with the NNK level (Figure 5). Since PON is considered as a potential NNK precursor, it is expected that this compound presents a positive correlation with total and matrix-bound NNK.[15]

The antioxidant properties of polyphenols may be able to explain the opposite behavior observed between polyphenols and oxidative products from nicotine. It is well-known that TSNAs are formed in tobacco leaves during the curing process by nitrosation of the alkaloids.[16] Therefore, it can be inferred that the polyphenols presented in the tobacco leaves act like an alkaloid protector, minimizing their nitrosation and, consequently, decreasing the concentration of TSNAs in cured tobacco. Likewise, a previous work showed that the polyphenols antioxidant capacity throughout the curing time of air-cured Burley reduced in 80%, during the first 2 weeks of curing, while the TSNAs and nitrite levels increased significantly after this period.[30] The major reason for the reduction of the antioxidant capacity was attributed to the decrease in the chlorogenic acid and rutin level after 2 weeks. Therefore, these findings help characterize the opposite behavior between the antioxidant capacity and the TSNAs content in cured tobacco.

3.4. HTS-FIA-HRMS Method Extension To Estimate Soluble Nitrosamines from Chemical Fingerprint.

From the proposed methodologies, HTS-FIA-HRMS was considered the most suitable method, allowing to predict not only matrix-bound NNK but also soluble and total NNK, regardless of the sample matrix or origin. It had errors within the expected range and further figures of merit to evaluate the model in accordance with the reported guideline.[31] Therefore, this methodology was applied to other soluble nitrosamines (NAB, NAT and NNN) determined through the GC-TEA reference method. The same chemometric parameters as for the NNK methods were used. The sample selection based on LOQ of the reference value, T^2, and DModX; bootstrapping of one-

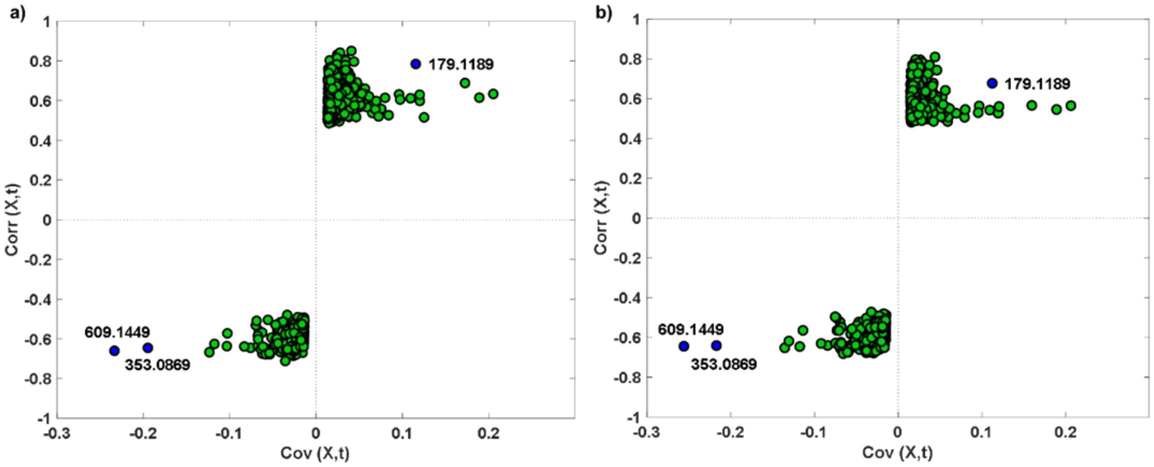


Figure 5. S-plot from O-PLS models of (a) matrix-bound NNK and (b) total NNK quantification.

third of the sample data set for validation with 100 repetitions. O-PLS models used Venetian blind with 7 windowed samples as cross-validation, VIP (threshold of 1), and correlation coefficient (threshold of 0.1) for the variable selection.

Table 2 presents the results obtained for O-PLS models of soluble NAB, NAT, and NNN, with their respective number of selected samples, minimal and maximum response values, R^2, R^2CV, R^2V, RMSEC, RMSECV, RMSEV, % error based on the RMSEV and the range of the model, LV, and number of selected variables. Due to the high limit of detection for the NAB reference method, only 88 samples were used to calibrate and validate this model. However, the bootstrapping methodology can still ensure model robustness and evaluate any possible overfitting by comparing the model evolution among the 100 repetitions. All models presented suitable results, with high coefficients of determination between 0.77 and 0.82 for cross-validation samples and 0.78 and 0.82 for validation samples, and close values for RMSEC, RMSECV and RMSEV among each response, demonstrating that no overfitting had occurred in the models. The relative errors were between 6.31 and 10.97%.

Table 2. Summarized Results for Quantitation of Soluble NAB, NAT, and NNN by HTS-FIA-HRMS and O-PLS Proposed Method

analyte	NAB	NAT	NNN
number of samples	88	471	328
min (ng·g^{-1})[a]	60.00	80.00	80.00
max (ng·g^{-1})[b]	406.67	2455.00	3593.33
R^2[c]	0.87	0.85	0.80
R^2CV[d]	0.80	0.82	0.77
R^2V[e]	0.82	0.81	0.78
RMSEC (ng·g^{-1})[f]	25.82	172.12	203.54
RMSECV (ng·g^{-1})[g]	32.51	187.57	218.80
RMSEV (ng·g^{-1})[h]	38.39	197.59	226.32
% error[i]	10.97	8.08	6.31
LV[j]	4	6	4
V_{sel}[k]	627	923	969

[a] Minimum value of concentration range; [b] Maximum value of concentration range. [c] Coefficient of determination of calibration values. [d] Coefficient of determination of cross-validation values. [e] Coefficient of determination of bootstrapping validation values. [f] Root-mean-square error of calibration. [g] Root-mean-square error of cross validation; [h] Root-mean-square error of bootstrapping validation. [i] Percentage error based on RMSEV and concentration range. [j] Number of latent variables. [k] Number of selected variables.

4. CONCLUSIONS

Three different approaches were tested to predict matrix-bound NNK. First, a linear regression methodology relating matrix-bound to soluble NNK. Then, a slightly more complex one, based on multiple linear regression, where several parameters were used to predict matrix-bound NNK. Finally, a method that uses HTS-FIA-HRMS and multivariate regression by orthogonal partial least-squares that permits the prediction of the two sources of NNK, soluble and matrix-bound, as well as total NNK.

The linear regression approach demonstrated a high influence of the matrix and country origin effect in a direct correlation between matrix-bound and soluble NNK, hindering the prediction performance of the models. The multiple linear regression approach, considering soluble NNK and other sample attributes, like color, ripeness, quality, amount of nicotine, and origin, showed a good correlation between the values predicted by the model and the reference values of matrix-bound NNK. An influence of the sample origin and nicotine amount was observed for both models, color being an important parameter for the flue-cured Virginia model. However, these results were limited to a previous knowledge of the sample and subjective sample parameters.

A novel approach through HTS-FIA-HRMS and multivariate regression by orthogonal partial least-squares was developed and validated with errors comparable to the reference method, but with a higher analytical capacity and predictability. This system permits to simultaneously predict the concentration of several TSNAs in one single analysis, in which soluble, matrix-bound, and total NNK as well as soluble NAT, NAB, and NNN, were verified. Considering all the models presented, the percentage of relative errors was between 5.25 and 11.98%. Although some models presented a limited number of samples, like the soluble NAB model, this approach easily permitted to expand the model by evaluating the samples with reported diagnostic tools (like DModX, T^2, standardized residual of response matrix) and a performing reference analysis.

In addition, the proposed methodology based on HTS-FIA-HRMS and O-PLS permits to analyze and process more than

700 samples of cured tobacco per day, with a simple extraction and experimental procedure and high throughput screening, resulting in a robust methodology without any chromatographic step. Therefore, this approach is an important tool for large tobacco monitoring programs with suitable accuracy and robustness.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.chemrestox.8b00154.

> Detailed coefficient values for MLR models, histograms of permutation tests for HTS-FIA-HRMS O-PLS models and control charts of the predicted values of matrix-bound NNK for control samples (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: gpsabin@gmail.com.

ORCID

Guilherme P. Sabin: 0000-0002-9353-9659

Notes

The authors declare no competing financial interest.

■ REFERENCES

(1) Bodhicharla, R., Ryde, I. T., Prasad, G. L., and Meyer, J. N. (2014) The tobacco-specific nitrosamine 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone (NNK) induces mitochondrial and nuclear DNA damage in Caenorhabditis elegans. *Environ. Mol. Mutagen. 55*, 43−50.

(2) Hecht, S. S. (1998) Biochemistry, Biology, and Carcinogenicity of Tobacco-Specific N-Nitrosamines. *Chem. Res. Toxicol. 11*, 559−603.

(3) Shi, H., Wang, R., Bush, L. P., Zhou, J., Yang, H., Fannin, N., and Bai, R. (2013) Changes in TSNA Contents during Tobacco Storage and the Effect of Temperature and Nitrate Level on TSNA Formation. *J. Agric. Food Chem. 61*, 11588−11594.

(4) Konstantinou, E., Fotopoulou, F., Drosos, A., Dimakopoulou, N., Zagoriti, Z., Niarchos, A., Makrynioti, D., Kouretas, D., Farsalinos, K., Lagoumintzis, G., and Poulas, K. (2018) Tobacco-specific nitrosamines: A literature review. *Food Chem. Toxicol. 118*, 198−203.

(5) Hoffmann, D., and Hecht, S. S. (1985) Nicotine-derived N-nitrosamines and tobacco-related cancer: current status and future directions. *Cancer Res. 45*, 935−944.

(6) (2007) *Smokeless Tobacco and Some Tobacco-Specific N-Nitrosamines; IARC Monographs on the Evaluation of Carcinogenic Risks to Humans*, Vol *89*, pp 9−626, International Agency for Research on Cancer, Lyon, France.

(7) (2004) *Tobacco Smoke and Involuntary Smoking; IARC monographs on the Evaluation of Carcinogenic Risks to Humans*, pp 1−1452, International Agency for Research on Cancer, Lyon, France.

(8) Edwards, S. H., Rossiter, L. M., Taylor, K. M., Holman, M. R., Zhang, L., Ding, Y. S., and Watson, C. H. (2017) Tobacco-Specific Nitrosamines in the Tobacco and Mainstream Smoke of U.S. Commercial Cigarettes. *Chem. Res. Toxicol. 30*, 540−551.

(9) Wittel, U. A., Momi, N., Seifert, G., Wiech, T., Hopt, U. T., and Batra, S. K. (2012) The pathobiological impact of cigarette smoke on pancreatic cancer development. *Int. J. Oncol. 41*, 5−14.

(10) Hecht, S. S., Rivenson, A., Braley, J., DiBello, J., Adams, J. D., and Hoffmann, D. (1986) Induction of Oral Cavity Tumors in F344 Rats by Tobacco-specific Nitrosamines and Snuff. *Cancer Res. 46*, 4162−4166.

(11) Hecht, S. S., Chen, C.-h. B., Ohmori, T., and Hoffmann, D. (1980) Comparative Carcinogenicity in F344 Rats of the Tobacco-specific Nitrosamines, *N*′-Nitrosonornicotine and 4-(*N*-Methyl-*N*-nitrosamino)-1-(3-pyridyl)-1-butanone. *Cancer Res. 40*, 298−302.

(12) Hoffmann, D., Rivenson, A., Amin, S., and Hecht, S. S. (1984) Dose-response study of the carcinogenicity of tobacco-specific N-nitrosamines in F344 rats. *J. Cancer Res. Clin. Oncol. 108*, 81−86.

(13) Hecht, S. S. (2003) Tobacco carcinogens, their biomarkers and tobacco-induced cancer. *Nat. Rev. Cancer 3*, 733−744.

(14) Gunduz, I., Kondylis, A., Jaccard, G., Renaud, J. M., Hofer, R., Ruffieux, L., and Gadani, F. (2016) Tobacco-specific N-nitrosamines NNN and NNK levels in cigarette brands between 2000 and 2014. *Regul. Toxicol. Pharmacol. 76*, 113−120.

(15) Lang, G., and Vuarnoz, A. (2015) Matrix-Bound 4-(Methylnitrosamino)-1-(3-pyridyl)-1-butanone in Tobacco: Quantification and Evidence for an Origin from Lignin-Incorporated Alkaloids. *J. Nat. Prod. 78*, 85−92.

(16) Lipowicz, P. J., and Seeman, J. I. (2017) A Model To Estimate the Sources of Tobacco-Specific Nitrosamines in Cigarette Smoke. *Chem. Res. Toxicol. 30*, 1556−1561.

(17) ISO/TS 22304:2008. (2008) *Tobacco - Determination of tobacco specific nitrosamines - Method using alkaline dichloromethane extraction*; International Organization for Standardization, Geneva.

(18) Pontes, O. S., Sabin, G. P., da Silva, J. R. P., Dias, J. C., and Kaiser, S. *Apparatus and method for classifying a tobacco sample into one of a predefined set of taste categories*. WO2018007789A1, 2018.

(19) Hotelling, H. (1933) Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol. 24*, 417−441.

(20) Wold, S., Kettaneh, N., Fridén, H., and Holmberg, A. (1998) Modelling and diagnostics of batch processes and analogous kinetic experiments. *Chemom. Intell. Lab. Syst. 44*, 331−340.

(21) Krishnan, A., Williams, L. J., McIntosh, A. R., and Abdi, H. (2011) Partial Least Squares (PLS) methods for neuroimaging: A tutorial and review. *NeuroImage 56*, 455−475.

(22) Hair, J. F., Ringle, C. M., and Sarstedt, M. (2011) PLS-SEM: Indeed a Silver Bullet. *J. Mark. Theory Pract. 19*, 139−152.

(23) Efron, B., and Tibshirani, R. (1986) Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. *Statist. Sci. 1*, 54−75.

(24) Leffingwell, J. C. (1999) Basic Constituents of Tobacco Leaf and Differences among Tobacco Types, In *Tobacco: Production Chemistry and Technology* (Davis, D. L., and Nielsen, M. T., Eds.) pp 265−284, Blackwell Science, Cambridge.

(25) Schröder, H. F. (1993) Surfactants: non-biodegradable, significant pollutants in sewage treatment plant effluents: Separation, identification and quantification by liquid chromatography, flow-injection analysis—mass spectrometry and tandem mass spectrometry. *J. Chromatogr. A 647*, 219−234.

(26) McCooeye, M., Ding, L., Gardner, G. J., Fraser, C. A., Lam, J., Sturgeon, R. E., and Mester, Z. (2003) Separation and Quantitation of the Stereoisomers of Ephedra Alkaloids in Natural Health Products Using Flow Injection-Electrospray Ionization-High Field Asymmetric Waveform Ion Mobility Spectrometry-Mass Spectrometry. *Anal. Chem. 75*, 2538−2542.

(27) Wang, W., Zhou, H., Lin, H., Roy, S., Shaler, T. A., Hill, L. R., Norton, S., Kumar, P., Anderle, M., and Becker, C. H. (2003) Quantification of Proteins and Metabolites by Mass Spectrometry without Isotopic Labeling or Spiked Standards. *Anal. Chem. 75*, 4818−4826.

(28) Everley, R. A., Kunz, R. C., McAllister, F. E., and Gygi, S. P. (2013) Increasing Throughput in Targeted Proteomics Assays: 54-Plex Quantitation in a Single Mass Spectrometry Run. *Anal. Chem. 85*, 5340−5346.

(29) Wiklund, S., Johansson, E., Sjöström, L., Mellerowicz, E. J., Edlund, U., Shockcor, J. P., Gottfries, J., Moritz, T., and Trygg, J. (2008) Visualization of GC/TOF-MS-Based Metabolomics Data for Identification of Biochemically Interesting Compounds Using OPLS Class Models. *Anal. Chem. 80*, 115−122.

(30) Krauss, M., Li, Q., and Hempfling, W. *Method for the reduction of tobacco specific nitrosamines by increasing antioxidants in tobacco*. US20030056801A1, 2003.

(31) (2017) *USP-40 General Information*, pp 963−981, United States Pharmacopeia, Rockville, MD.