Nitrate and nitrite as mixed source of nitrogen for Chlorella vulgaris: fast nitrogen quantification using spectrophotometer and machine learning

Victor Pozzobon, Wendie Levasseur, Cédric Guerin, Patrick Perré

▶ **To cite this version:**

Victor Pozzobon, Wendie Levasseur, Cédric Guerin, Patrick Perré. Nitrate and nitrite as mixed source of nitrogen for Chlorella vulgaris: fast nitrogen quantification using spectrophotometer and machine learning. Journal of Applied Phycology, 2021, 33 (3), pp.1389-1397. 10.1007/s10811-021-02422-2 . hal-03384999

HAL Id: hal-03384999
https://hal.science/hal-03384999v1

Submitted on 19 Oct 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Nitrate and nitrite as mixed source of nitrogen for *Chlorella vulgaris*: fast nitrogen quantification using spectrophotometer and machine learning

Victor Pozzobon[1] ✉, **Wendie Levasseur**[1], **Cédric Guérin**[1], **and Patrick Perré**[1]

[1]LGPM, CentraleSupélec, Université Paris-Saclay, SFR Condorcet FR CNRS 3417, Centre Européen de Biotechnologie et de Bioéconomie (CEBB), 3 rue des Rouges Terres 51110 Pomacle, France

This article presents a machine learning workflow allowing to construct spectrophotometric equations predicting nitrate and nitrite concentrations within microalgae culture samples. First, numerous samples with various nitrate and nitrite concentrations (in mixture or separated) were drawn from cultures. Their UV absorbance spectra were recorded with a tabletop spectrophotometer before being analyzed using ion chromatography. Then, the data collected were used to construct a machine leaning model based on partial least square regression. From a practical perspective, the best model involves 3 wavelengths to quantify both nitrate and nitrite within the samples. The proposed equations can readily be used (LoQ of 0.5 mg.L^{-1}, uncertainty of $\pm$ 10 %). They greatly shorten the delay to obtain sample nitrate and nitrite concentrations compared to ion chromatography while retaining adequate accuracy. Furthermore, the workflow is presented step-wisely, with emphasis on relevant details so that other scholars may deploy in their own laboratory to best suit their own needs. Finally, the data and source files are made available in an online repository.

Nitrogen | Quantification | Spectrophotometry | Machine learning | Partial least square

Correspondence: ***victor.pozzobon@centralesupelec.fr***

1. Introduction

Efficient microalgal cultivation requires to supply the cells with adequate nutrients in the proper amount within a growth promoting environment. Those nutrients include macronutrients, such as carbon and nitrogen, and micronutrients, such as trace elements and vitamins. Not only one has to supply them adequately at the beginning of the culture, but one should also ensure their availability throughout the whole process or control their depletion if a stressing strategy is intended. The latter strategies are quite common in microalgal biotechnology as they can trigger secondary metabolites hyperaccumulation (1, 2). The most common ones are salinity stress, light stress and nitrogen stress.

Nitrogen is an essential macronutrient that takes part in amino, proteins, enzymes and nucleic acids syntheses. It can be delivered to microalgae under three inorganic forms (ammonium, nitrate and nitrite) or under various organic forms (the most common being urea) (3). Among the difference sources of nitrogen, nitrate is the widest spread one within laboratories as it comes into play in almost every synthetic medium (4). Urea and nitrite are more commonly encountered in wastewater treatment applications. Ammonium, while it is microalgae favorite source of nitrogen, is rarely used from a biotechnological perspective because it is highly volatile.

Nitrate key role in the culture of microalgae makes it a vital parameter to follow. To do so, several techniques have been deployed over the years. They range from basic colorimetric approaches to advanced chromatographic methods. The first ones are relatively inexpensive but may require to manipulate hazardous chemical species, or simply involves numerous sample manipulations, making them undesirable (5–7). The second ones depend upon high end equipment requiring sizable capital expenditure and expertise (8). In addition, none of these techniques yield immediate results and both lead to sample destruction. Still, when operating a microalgal process, swift actions may sometimes be required not to lose the culture. A third kind of methods can yield nitrate quantification almost instantaneously, maybe at the price of a somewhat lower accuracy, namely UV spectrophotometric methods.

UV spectrophotometric nitrate and nitrite quantifications are known to the ecotoxicity community but are singularly unfamiliar to microalgal biotechnologists. This is all the more surprising as, for example, spectrophotometric pigments quantification is a basic laboratory procedure (9). For nitrate and nitrite, the working principle is the same as for pigments. Peaks absorbance wavelengths are used to identify each species contribution, and a wavelength far from those peaks is used to nullify background noise originating from dissolved organic matter (10–13). Resulting equations (also sometimes referred to as algorithms) are calibrated either using artificial or genuine samples whose absorbance spectra are correlated with chromatographic measurements. In the case of nitrate and nitrite, this procedure is greatly facilitated by the existence of very strong absorbance peaks in the UV range (190 - 230 nm), in addition to moderate ones in the lower part (275 - 400 nm) of the visible spectrum (Fig. 1). This represents an advantage. Indeed, in order to access UV absorbance peaks associated with nitrate and nitrite (190 - 230 nm), one has to considerably dilute the sample (more

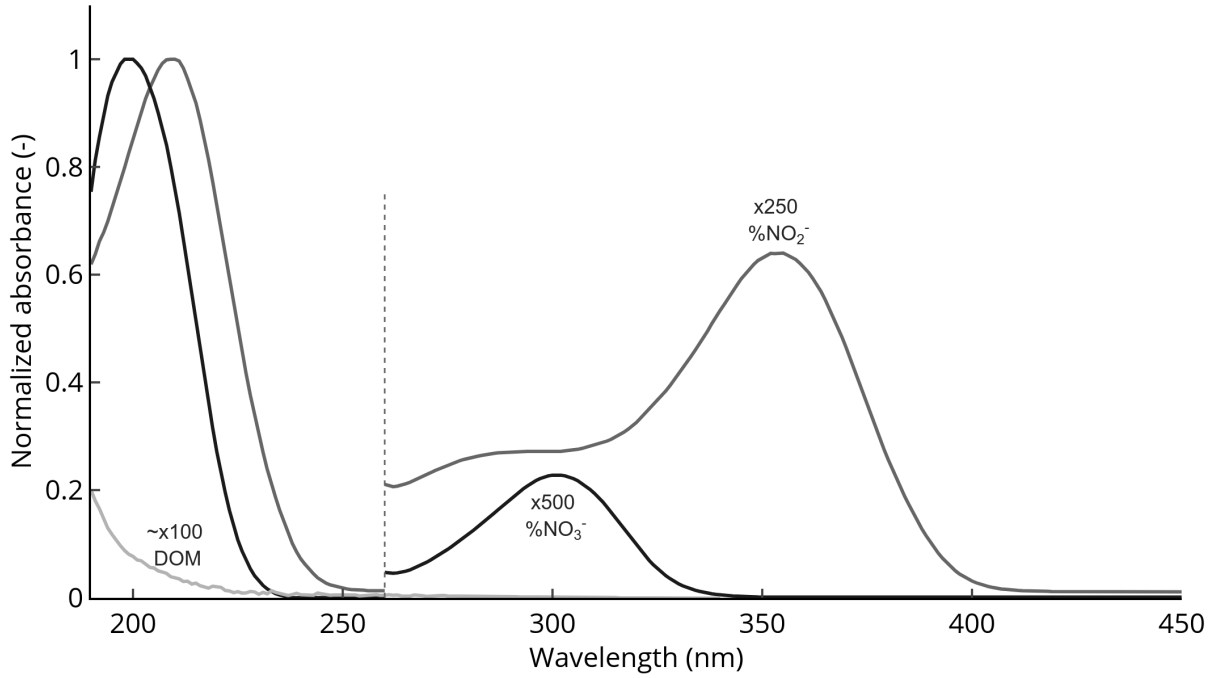

Fig. 1. Normalized absorbance spectra of pure nitrate, pure nitrite and Dissolved Organic Matter (DOM). Obtained with two measurements with two different concentrations for each species. Black line: nitrate, gray line: nitrite, light gray: Dissolved Organic Matter (measured from 190 to 340 nm after a special run of two weeks including 1 week nitrogen source exhaustion)

than 200 fold in our case). This procedure almost nullifying background noise originating from dissolved organic matter.

While efficient and easy to deploy, current spectrophotometric nitrate and nitrite quantifications can be regarded as suboptimal in two ways. First, wavelengths corresponding to chemical species are selected beforehand and may not account for potential matrix effect. Then, dissolved organic matter contribution may vary for sampling location to sampling location. Luckily, this second drawback is unlikely to be encountered in biotechnology laboratories, where cultures are well controlled and usually carried out on a time scale short enough to prevent excessive dissolved organic matter accumulation. Still, the question of wavelengths selection remains.

Using tabletop spectrophotometers to acquire absorbance spectra of dissolved molecules mixtures to quantify them is a well-established technique. A whole field of research has been dedicated to this question over many years and produced streamlined workflows and mathematical background to support this technique. In a nutshell, mixtures with known concentrations of the species of interest are created and the corresponding spectrum is acquired. This constitutes a dataset that is then used to calibrate a model. This model correlates input spectra with output concentrations. Among the candidate models, such as Principal Component Analysis or MultiLinear Regression, Partial Least Square (or PLS) regression algorithm is of choice (14) and has also proven successful in microalgal biotechnology applications (15, 16). Its particularity is that it uses principal component decompositions to create a set of components (linear combination of variables) associated to both input and output variables (through maximization of covariance between the scores). In this way, the most meaningful information is retained, making it a robust model (low sensibility to the training data) that handles well colinear inputs (when multiple variables provide the same information, two neighboring wavelengths in the case of a spectrum) (17). Still, in the case of spectrophotometric readings processing, two metaparameters remain to be optimized by the operator: the number of components (boiling down to the number of wavelengths that intervene in the correlations) and the selected wavelengths themselves.

This article reports how such technique can be deployed within a biotechnology laboratory to produce a correlation linking spectrophotometric measurements to nitrate and nitrite concentrations. Spectra and nitrogen sources concentrations were obtained from *Chlorella vulgaris* cultures. It yielded a rich dataset over a wide range of nitrate and nitrite concentrations at different stages of the culture. They were then used to power the machine learning workfow processing quantifying nitrate and nitrite concentrations from those spectra. Finally, the source files associated to this work are freely available in an online repository for anyone to download. This way the interested reader could deploy the workflow and obtain correlations best-suiting its need (nitrogen sources, type of biomass, ...).

2. Experiments and data acquisition

2 1. Culture protocol

Chlorella vulgaris (CV 211-11b) obtained from SAG Culture Collection, Germany, were cultivated on B3N medium variants (4). Seven alternative media were formulated with the same total molar nitrogen content, only the source was varied, either nitrate (from $NaNO_3$) or nitrite (from $NaNO_2$).

The tested nitrite fractions were: 0, 20, 40, 50, 60, 80 and 100 % NO_2^-. Cultures were carried out over seven days, samples being drawn twice a day for both cell growth monitoring and nitrogen quantification. The study was carried out in biological triplicates. A total of 273 samples were produced during this experimental campaign.

2 2. Nitrate and nitrite concentrations monitoring and spectra recording

Nitrate and nitrite concentrations were determined for all the sampling points. Samples were prepared by filtration (0.22 µm polypropylene) before being diluted with milliQ water (Integral-5, Merck, also analysed for correction) in order to be in the range of 0.2 to 10 mg.L^{-1}, corresponding to a peak absorbance of about 0.5. UV absorption spectrum (190 - 340 nm) was acquired using a spectrophotometer (Shimadzu UV-1800). Nitrate and nitrite quantifications were carried out on an ICS-5000+ Ion Chromatography system (Thermo Fisher Scientific) coupled with a conductivity detector. Separation was achieved on an AS11-HC column (2x250 mm, 4 µm) protected by a guard column AG11-HC (2x50 mm, 4 µm). Column temperature was maintained at 35 °C. The mobile phase was potassium hydroxide at a flow rate of 0.3 mL.min^{-1} and was delivered by an EGC 500 eluent generator. Elution was achieved in gradient mode; initial concentration of KOH was 5 mM which was held for 0.5 minute, the concentration was gradually increased to 25 mM in 15.5 minutes, then rapidly to 30 mM in 0.1 minute and held at 350 mM for 8 minutes followed by going back to initial composition and stabilization of the column for 10 more minutes. The eluent generated was then purified by a CR-TC to trap impurities such as carbonate. The suppressor was an ADRS 600, 4 mm, operated in recycle mode at 38 mA. The injection volume was 2 µl and total run analysis was 34 minutes. All ions were identified by comparison with their retention times with standard solutions. Quantification was achieved using the height of the peak in external calibration, the range of concentrations was from 0.2 to 10 mg.L^{-1}. All standards were purchased from Sigma-Aldrich with a TraceCert quality which is a standard at 1000 $\pm$ 4 mg.L^{-1}. This protocol ensures that the nitrate and nitrite concentrations can be directly related to the measured spectra as there was no intermediary manipulations between spectra acquisition and ion chromatography quantification.

3. Data management

The experimental campaign resulted in 273 data points: absorbance spectrum from 190 to 340 nm, resolution 1 nm resulting in 151 input variables, and measurements of nitrate and nitrite concentrations, resulting in 2 output variables. As a first step, manual curation of the data was undergone. 6 data points were excluded as they exhibited distorted spectra. 6 others data points were discarded because of obviously incorrect or lacking concentration measurements. The second step was to transform data so that all the values could be processed by PLS algorithm. Indeed, nitrate or nitrite values below the limit of quantification could be reported either as 'N.A.' or a random value below 0.2 mg.mL^{-1}. The question of data replacement in the case of a value below detection limit for PLS algorithm was already investigated in depth by other scholars (18). Their conclusions are clear when the actual value is known, one can use it to replace the machine reading. Otherwise, replacing the value by 0 is a safe procedure as it does not induce a bias and limits variance. In our case, thanks to our experimental design, we have access to the known value, which is 0. Thus, values below the detection limit were replaced by 0.

Data were then split into two datasets, one for calibration (80 % of the total, randomly drawn, n = 208) and one for validation (complementary 20 %, n = 53). The validation dataset is put aside until the very last stage of the protocol. The calibration dataset was further split into a training dataset (80 % of the calibration dataset, randomly drawn, n = 166) and a test dataset (complementary 20 %, n = 42). Training and test datasets were used to identify the relevant wavelengths to include in the PLS algorithm. Once the wavelengths were selected, the algorithm was calibrated on the calibration dataset as a whole. Finally, the algorithm predictions were tested on the validation dataset. This way the quality of the prediction is assessed from data never presented to the algorithm before. For the sake of readability, the data management workflow is summarized in Figure 2.

4. Partial least square calibration

As presented in the introduction, in our case, the partial least square model features two metaparameters to optimize: the number of components, in our case simply the number of wavelengths to be used to predict the concentrations (e.g. 3, 4, ... wavelengths in total), and the list of the particular wavelengths that are retained (e.g. 200 nm, 232 nm, ...).

4 1. Optimal number of components

The first step was to determine the optimal number of wavelengths to include in the final correlation. This number is a tradeoff between the improvement each additional wavelength adds to the model and the actual usability of the model. Regarding usability, 3-wavelength spectrophotometric correlations are common and easy to implement in tabletop spectrophotometer user interface. On the contrary, one would easily understand that a 9-wavelength correlation quantifying two concentrations would be quite troublesome to use and raise doubts about the relevance of all the involved wavelengths. Furthermore, from a technical perspective, using an excessive number of wavelengths would lead to overfitting, *i.e.* very good performances on the calibration dataset which are not reproducible.

Determining which set of wavelengths is relevant beforehand is not an easy task (19). The first possibility is to let the human operator select the wavelengths. An adequate starting point would be to retain one for each absorption peak of the species (nitrate, nitrite, dissolved organic matter). Still, the question of how to select additional wavelengths remains. Another possibility is to use a numerical optimizer to selected

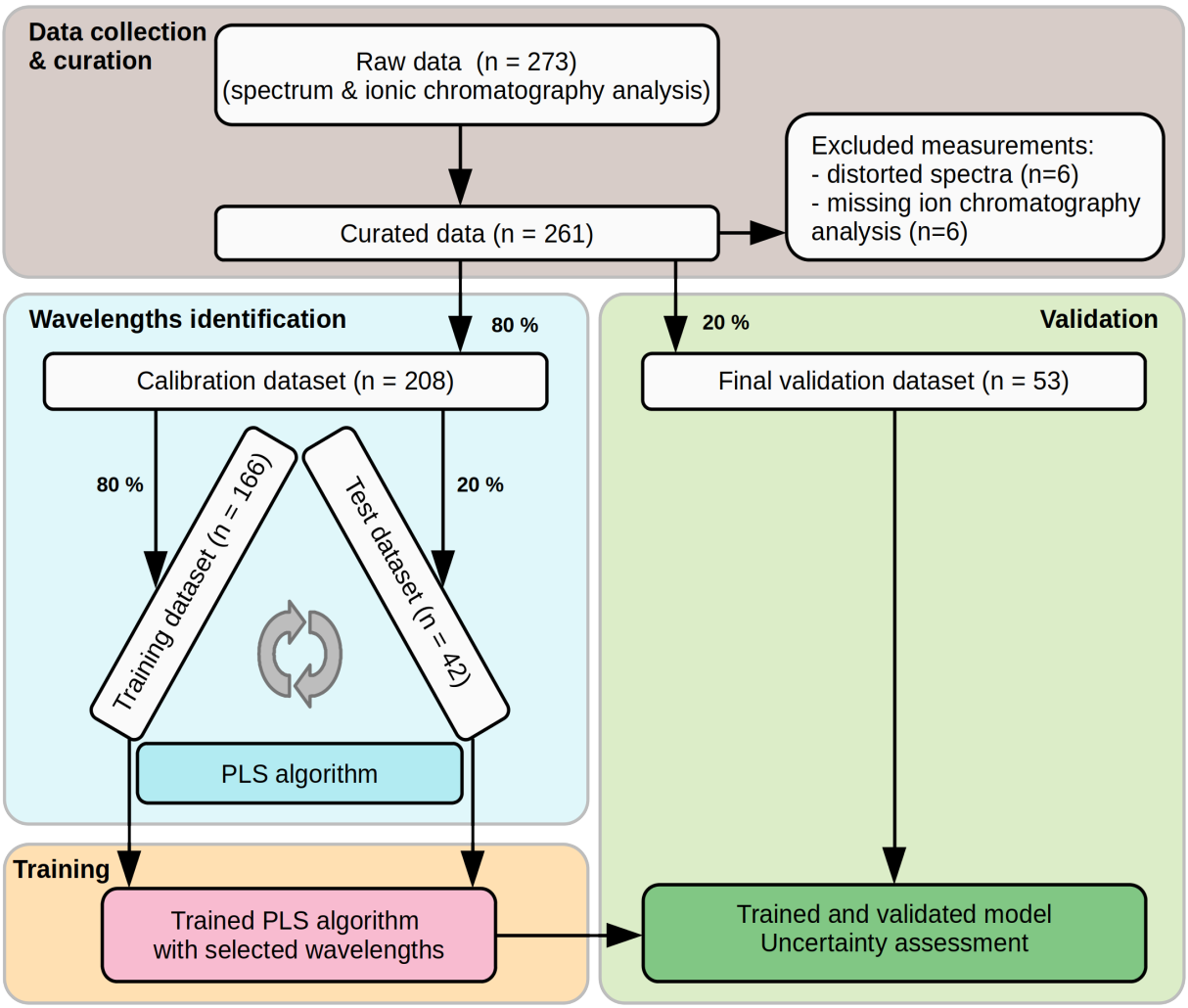

Fig. 2. Model development workflow and data management schematic

the most suitable wavelengths. We chose the second possibility.

The numerical procedure is as follows:

- First, the optimizer selected a subset of the 151 possible wavelengths, trains the PLS algorithm on training dataset and evaluates its performances on the test dataset. From this, it produces a cost function value associated to this first subset of wavelengths,
- Then, it determines another subset of wavelengths likely to produce a lower cost function value and evaluate it

Still, a special care has to be taken in choosing the cost function associated to this optimizer. Indeed, it is very likely that including a large number of wavelengths would produce lower cost function values biasing the process towards correlations featuring a large number of wavelengths. Though the question of the information added by each newly included wavelength is to be raised. Luckily, a special metric, called the Akaike Information Criterion (Eq. 1) (20), has been engineered to deal with this problem. This metric combines both the residual sum of square (RSS, over n data points), meaning the prediction error, and the number of wavelengths involved (k) to achieve this prediction. This way, models with a high number of wavelengths but rather poor prediction improvement are discarded.

$$AIC = 2k + n\ln(RSS) \quad \textbf{(1)}$$

Then, the question of the choice of the optimizer itself has to be addressed. Given the very high number of possible combinations (2^{151}), brute force approach lies out of the scope. The inclusion, or not, of a wavelength being a boolean value and considering the large dimensionality (151 variables) of the problem, gradient based methods do not seem appropriate either. Stochastic methods on the contrary have been shown to cope well with such configurations. Among them, particle swarm optimization is of note (21), as it is rather easy to implement, to deploy on a parallel architecture and capable of browsing considerable candidate spaces. Still, its main drawback is that because of the social nature of the swarm, it can sometimes lead to premature convergence. To avoid this pitfall, it can be coupled with another stochastic optimization method: the genetic algorithms. Together, these optimization techniques form an adequate tool to solve the problem at hands (for more details, the reader can refer

to (22), a practical implementation can be found in the repository associated to this article). The optimization algorithm parameters were set as follows: 400 particles/exemplars, cognitive and social parameters of the swarm were both set to 0.6, the inertia followed a random chaotic model, the mutation probability was set to 0.001. Runs were stopped after the swarm's best particle stagnated for 50 iterations.

Because of the stochastic nature of the optimizer and the considerable size of candidate space, one cannot be sure of uniqueness of the obtained solution. Thus, the optimizer was run 1000 times (lasting about 2 days and a half, using a laptop - Intel(R) Core(TM) i9-9880H CPU @ 2.30GHz -). Most of the time, it selected 3 wavelengths, rarely 4. Using those 1000 thousand runs, 151 wavelengths were ranked by occurrence. Then, models with increasing numbers of wavelengths, starting by the most reported ones, were trained and tested. The quality of the prediction was evaluated using Predicted Residual Error Sum of Squares (PRESS) as output metric. The influence of the number of wavelengths included on the prediction PRESS score can be found in Figure 3. In addition, the figure also reports the improvement associated with the step-wise wavelength addition. The first comment is that going from 1 to 2 wavelengths dramatically improves the prediction (88 %). This is normal as one wavelength alone cannot differentiate both nitrate and nitrite. Going to 3 and 4 wavelengths improved the prediction by 14 and 10 % respectively. Afterwards, any additional wavelength does not bring any relevant information (improvement $\pm$ 2 % centered around 0).

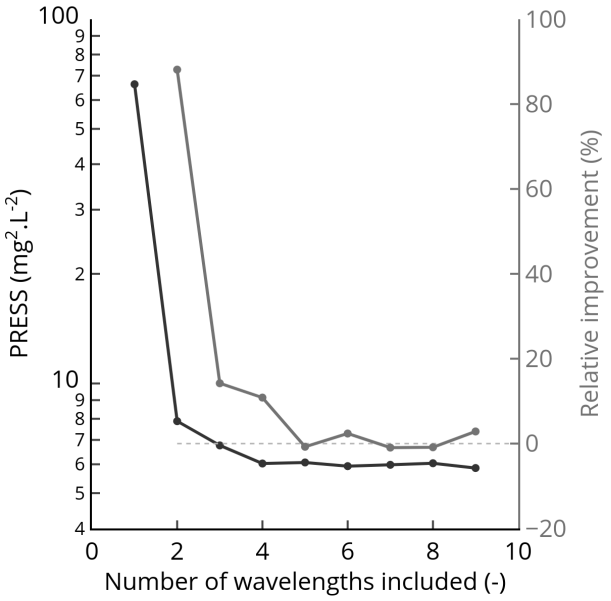

Fig. 3. PLS regression accuracy with increasing number of wavelengths included in the model. Black line: PRESS score. Grey line: improvement associated with the last wavelength addition

From this, it can be concluded that at least two wavelengths have to be included in the model, which is not surprising. In addition, depending of the complexity one wants to manipulate, 3 to 4 wavelengths in total could be included in the final model. In the remaining part of this work, 3 and 4-wavelength models are developed.

4 2. 3 and 4-wavelength models

The two models were calibrated separately using the same workflow, only the computational time changed. In this case, a systematic approach was chosen. This represents $151 \times 150 \times 149 = 3\ 374\ 850$ possible combinations for the 3-wavelength case and about 500 million for the 4-wavelength case. All of those combinations were tested. This time, the metric used to assess the quality of the prediction was the Mean Squared Error (MSE) as it is less computational intensive than the PRESS metric.

In the case of the 3-wavelength model, the retained wavelengths are: 200 nm, 224 nm and 234 nm (Eq. 2 and 3). The obtained Mean Squared Error on the test dataset is 38.8 10^{-3} mg^2.L^{-2}. By superimposing them on the species spectra (Fig. 4), one can gain some insights on the algorithm underlying logic. In this case, the first wavelength corresponds to the absorbance peak of nitrite. The second accounts for a combination of both nitrate and nitrite and the third one to nitrate only. All those wavelengths are located in the part of the spectrum where the absorbance is maximum (below 260 nm). Species by species now, given the values of coefficients, the main part of the nitrite concentration prediction relies on the difference between the two last wavelengths and seems to use the first as a correction (negative coefficient). For nitrate, explanation does not appear to be as straightforward as all the coefficients are of the same order of magnitude.

$$NO_2^- = -5.86A_{200nm} + 33.20A_{224nm} - 29.32A_{234nm} - 0.07 \quad \textbf{(2)}$$

$$NO_3^- = 8.42A_{200nm} - 9.95A_{224nm} - 7.69A_{234nm} - 0.10 \quad \textbf{(3)}$$

In the case of the 4-wavelength model, the retained wavelengths are: 208 nm, 210, nm, 312 nm and 334 nm (Eq. 4 and 5), and the Mean Squared Error is 35.0 10^{-3} mg^2.L^{-2}. The first comment is that the relative improvement actually associated with the addition of a fourth wavelength is of 9.8 % which is in agreement with previous section analysis. Then, here again, some insights can be gained by analyzing the selected wavelengths (Fig. 4). The two first wavelengths are close-by located and associated with the nitrate peak of highest intensity. The third seems to account for nitrite lowest intensity peak and the fourth to nitrate signal when nitrite one extinguishes. It is interesting to note that the two first wavelengths are so close that the nitrate absorbance does not change between the two while nitrite one varies a lot. This could allow the algorithm to fine tune its prediction of nitrite concentration. Regarding nitrate prediction, the constant term (-0.31 mg.L^{-1}) is relatively high and lies above ion chromatography limit of quantification (0.2 mg.L^{-1}). This remark points towards possible overfitting of the model. Still,

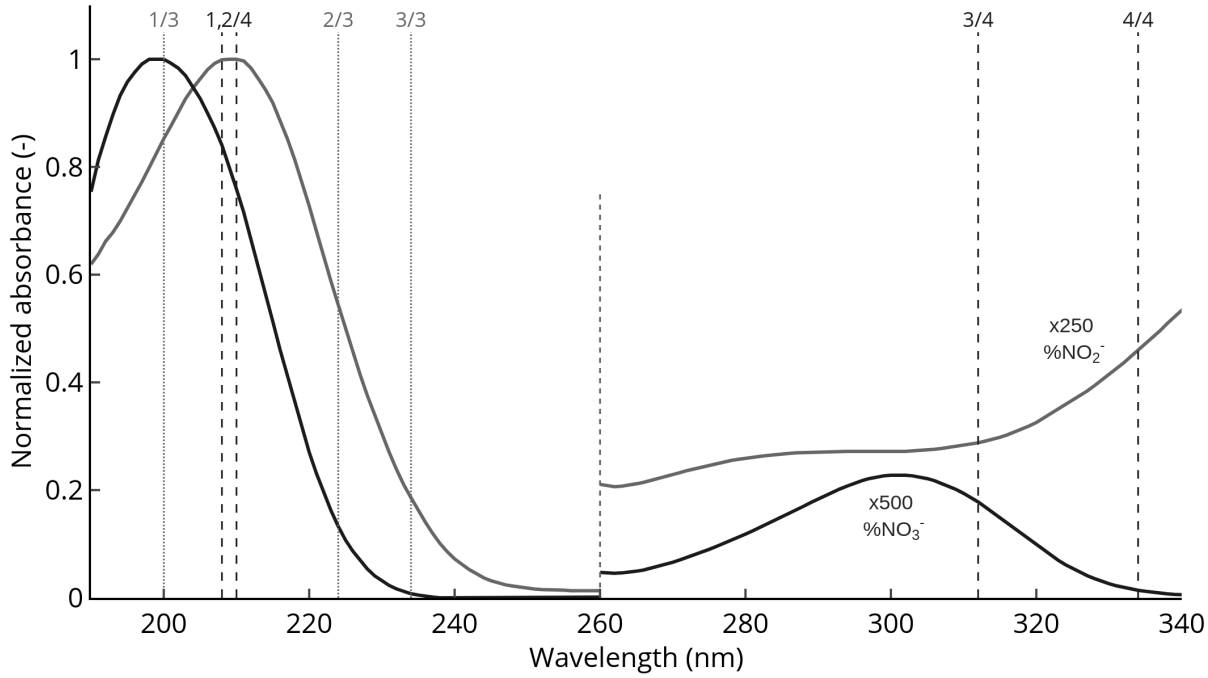

Fig. 4. Normalized absorbance spectra of pure nitrate, pure nitrite and retained wavelengths. Dotted gray lines: wavelengths retained for the 3-wavelength model. Dashed black: wavelengths retained for 4-wavelength model. Continuous black line: nitrate absorbance spectrum. Continuous gray line: nitrite absorbance spectrum

further dissection seems irrelevant as increasing the number of wavelengths increased the complexity of the equations lessening their explanability.

$$NO_2^- = -89.51A_{208nm} + 96.49A_{210nm} - 52.70A_{312nm} + 47.54A_{334nm} + 0.05 \quad \textbf{(4)}$$

$$NO_3^- = 83.06A_{208nm} - 81.77A_{210nm} - 97.16A_{312nm} + 90.53A_{334nm} - 0.31 \quad \textbf{(5)}$$

5. Algorithm validation

The next step after calibration was to challenge the models on a part of the dataset they never encountered before: the validation dataset. For the sake of readability, the 3-wavelength model is detailed first, then the differences with the 4-wavelength model are highlighted before drawing a recommendation on which to use.

Figures 5 and 6 report the comparison of the predicted and measured concentrations on the validation dataset for the 3-wavelength model. As one can see, most of the predictions fall within a $\pm$ 10 % interval around the measured value, few within a $\pm$ 20 % interval and almost none lying more than 20 % away of the measured value. Furthermore the spread around the first bisector is constant. In addition to being a token of the quality of the model, this means that the proposed model is capable of dealing indifferently with samples containing only nitrate, only nitrite or a mixture. These can be considered as very satisfactory results.

In order to dive further into the results, errors can be analyzed. Error distributions for nitrate and nitrite concentrations predictions are drawn on Figures 7 and 8. The errors are normally distributed around 0.000 mg.L^{-1} for nitrate and 0.011 mg.L^{-1} for nitrite, with standard deviations of 0.104 and 0.092 mg.L^{-1} respectively. Thus the proposed model exhibits no bias and a narrow spread of errors. Finally, Limit of Detection (LoD, Eq. 6) and Limit of Quantification (LoQ, Eq. 7) can be computed using mean blank value ($\bar{X}_b$) and blank standard deviation (σ_b) (23). As $\bar{X}_b$ turned out to be negative for both species (-0.074 and -0.021 mg.L^{-1} for nitrate and nitrite respectively), 0 was retained

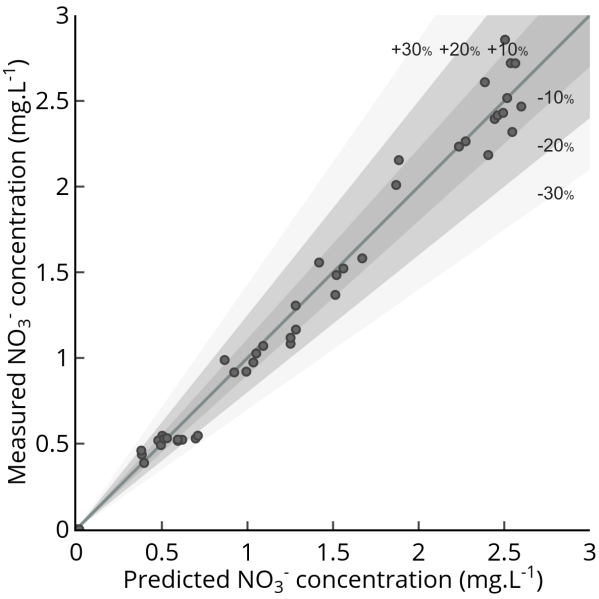

Fig. 5. Comparison of the predicted and measured nitrate concentrations for the 3-wavelength model on the validation dataset. Line: first bisector. Shaded areas: $\pm$10, $\pm$20 and $\pm$30 % deviation.

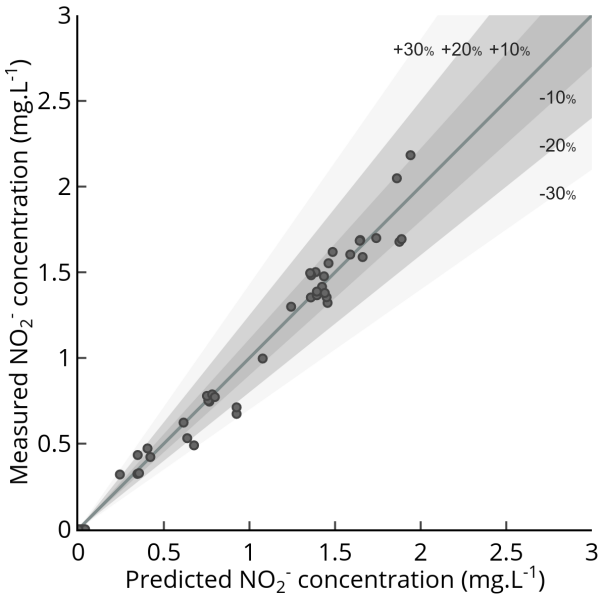

Fig. 6. Comparison of the predicted and measured nitrite concentrations for the 3-wavelength model on the validation dataset. Line: first bisector. Shaded areas: ±10, ±20 and ±30 % deviation.

instead. For nitrate, this procedure yielded a $\text{LoD}_{NO_3^-}$ of 0.153 mg.L^{-1} and a $\text{LoQ}_{NO_3^-}$ of 0.509 mg.L^{-1}. While for nitrite a $\text{LoD}_{NO_2^-}$ of 0.133 mg.L^{-1} and a $\text{LoQ}_{NO_2^-}$ of 0.444 mg.L^{-1} were obtained. These values are close to the one of ion chromatography. Furthermore, LoQs lie just below the lower bounds of the ranges of values used to calibrate the algorithm, which is normal.

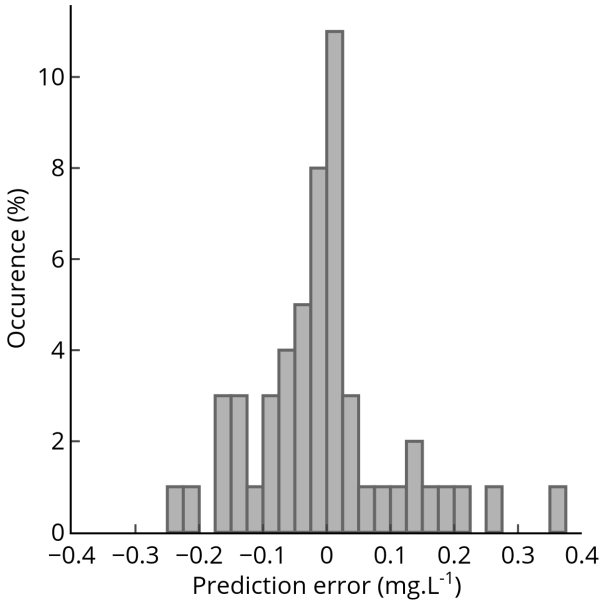

Fig. 7. Error distribution for the 3-wavelength model nitrate concentration prediction on the validation dataset. Bin width: 0.025 mg.L^{-1}

$$LoD = \bar{X_b} + 3\sigma_b \quad \textbf{(6)}$$

$$LoQ = \bar{X_b} + 10\sigma_b \quad \textbf{(7)}$$

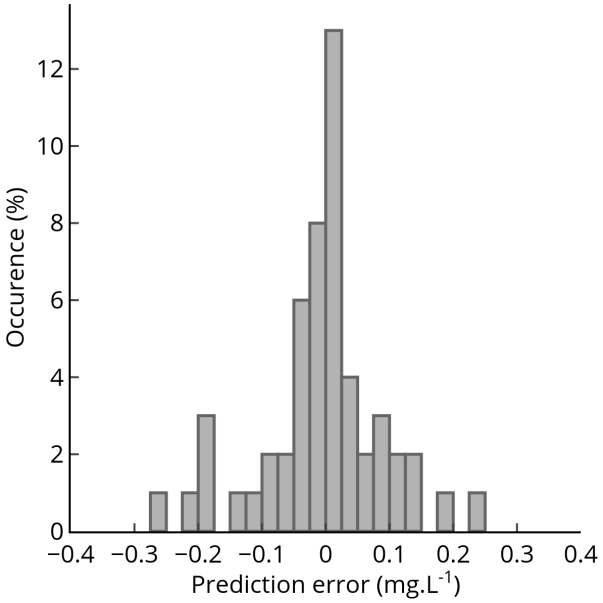

Fig. 8. Error distribution for the 3-wavelength model nitrite concentration prediction on the validation dataset. Bin width: 0.025 mg.L^{-1}

The same analyzes were undergone for the 4-wavelength model. While it provided results similar to the ones of the 3-wavelength model, differences arose. For example, the Mean Squared Error on the validation dataset was higher for the 4-wavelength (8.6 10^{-3} mg^2.L^{-2}) than for the 3-wavelength one (11.9 10^{-3} mg^2.L^{-2}). The error distribution standard deviations are also somewhat higher for the 4-wavelength model, 0.110 and 0.107 mg.L^{-1} for nitrate and nitrite respectively. It means that not only the predictions of this model are less accurate, they are also on average further away from the actual value than those of the 3-wavelength model. Recalling the high value of the constant term (-0.31 mg.L^{-1}) for the 4-wavelength model, these three observations point towards potential overfitting (high performances and dependence on the calibration dataset). Thus, we can only advise to use the 3-wavelength model, which is also incidentally more convenient to use.

6. Conclusion

This article presented a machine learning workflow allowing to construct spectrophotometric equations in order to quantify nitrate and nitrite within a sample. The quantification is based on three wavelengths: 200, 224 and 234 nm. From a practical perspective, the proposed model is not only calibrated but also carefully validated, so that the equations can readily be used (LoQ of 0.5 mg.L^{-1}, uncertainty of ± 10 %). This would greatly shorten the delay to obtain samples nitrate and nitrite concentrations (or only one of them) compared to ion chromatography while retaining adequate accuracy. Furthermore, the workflow is presented step-wisely, with emphasis on relevant details so that other scholars may deploy in their own laboratory. Finally, the data and source files are made available in an online repository.

ACKNOWLEDGEMENTS

This study was carried out in the Centre Européen de Biotechnologie et de Bioéconomie (CEBB), supported by Région Grand Est, Département de la Marne, Greater Reims and the European Union. In particular, the authors would like to thank Département de la Marne, Greater Reims, Région Grand Est and European Union with European Regional Development Fund (ERDF Champagne Ardenne 2014-2020) for their financial support to the Chair of Biotechnology of CentraleSupélec.

PLS implementation was directly drawn from Scikit Learn 0.23.2 machine learning framework (24).

AUTHOR CONTRIBUTIONS

WL and VP initiated and designed the study. AM and WL led the experimental work with the help of CG and VP. VP led the numerical work. All the authors critically interpreted the results. VP drafted the manuscript, the other authors corrected it. All authors approve the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Bibliography

1. Xiao-Man Sun, Lu-Jing Ren, Quan-Yu Zhao, Xiao-Jun Ji, and He Huang. Microalgae for the production of lipid and carotenoids: a review with focus on stress regulation and adaptation. *Biotechnology for Biofuels*, 11(1):272, October 2018. ISSN 1754-6834. .
2. Shih-Hsin Ho, Xiaoting Ye, Tomohisa Hasunuma, Jo-Shu Chang, and Akihiko Kondo. Perspectives on engineering strategies for improving biofuel production from microalgae — A critical review. *Biotechnology Advances*, 32(8):1448–1459, December 2014. ISSN 0734-9750. .
3. Johan A. Hellebust and Iftikhar Ahmad. Regulation of Nitrogen Assimilation in Green Microalgae. *Biological Oceanography*, 6(3-4):241–255, January 1989. ISSN 0196-5581. . Publisher: Taylor & Francis _eprint: https://www.tandfonline.com/doi/pdf/10.1080/01965581.1988.10749529.
4. Robert A. Andersen. Algal Culturing Techniques Appendix A—Recipes for Freshwater and Seawater Media. In *Algal Culturing Techniques*. Academic Press, Burlington, Mass, 1 edition edition, February 2005. ISBN 978-0-12-088426-1.
5. M. J. Taras. Phenoldisulfonic Acid Method of Determining Nitrate in Water. Photometric Study. *Analytical Chemistry*, 22(8):1020–1022, August 1950. ISSN 0003-2700. .
6. Katrina M. Miranda, Michael G. Espey, and David A. Wink. A Rapid, Simple Spectrophotometric Method for Simultaneous Detection of Nitrate and Nitrite. *Nitric Oxide*, 5(1):62–71, February 2001. ISSN 1089-8603. .
7. Nidal A Zatar, Maher A Abu-Eid, and Abdullah F Eid. Spectrophotometric determination of nitrite and nitrate using phosphomolybdenum blue complex. *Talanta*, 50(4):819–826, November 1999. ISSN 0039-9140. .
8. J. R. Thayer and R. C. Huffaker. Determination of nitrate and nitrite by high-pressure liquid chromatography: Comparison with other methods for nitrate determination. *Analytical Biochemistry*, 102(1):110–119, February 1980. ISSN 0003-2697. .
9. Alan R. Wellburn. The Spectral Determination of Chlorophylls a and b, as well as Total Carotenoids, Using Various Solvents with Spectrophotometers of Different Resolution. *Journal of Plant Physiology*, 144(3):307–313, September 1994. ISSN 0176-1617. .
10. Claudio Minero, Vittorio Lauri, Gianpaolo Falletti, Valter Maurino, Ezio Pelizzetti, and Davide Vione. Spectrophotometric characterisation of surface lakewater samples: implications for the quantification of nitrate and the properties of dissolved organic matter. *Annali di Chimica: Journal of Analytical, Environmental and Cultural Heritage Chemistry*, 97(10):1107–1116, 2007. ISSN 0003-4592. .
11. A. P. Carvalho, L. A. Meireles, and F. X. Malcata. Rapid spectrophotometric determination of nitrates and nitrites in marine aqueous culture media. *Analusis*, 26(9):347–351, November 1998. ISSN 0365-4877, 1286-482X. . Publisher: EDP Sciences.
12. Miles S. Finch, David J. Hydes, Charles H. Clayson, Bernhard Weigl, John Dakin, and Pat Gwilliam. A low power ultra violet spectrophotometer for measurement of nitrate in seawater: introduction, calibration and initial sea trials. *Analytica Chimica Acta*, 377(2): 167–177, December 1998. ISSN 0003-2670. .
13. Y. Collos, F. Mornet, A. Sciandra, N. Waser, A. Larson, and P.J. Harrison. An optical method for the rapid measurement of micromolar concentrations of nitrate in marine phytoplankton cultures. *Journal of Applied Phycology*, 11(2):179–184, April 1999. ISSN 1573-5176. .
14. Svante Wold, Michael Sjöström, and Lennart Eriksson. PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58(2):109–130, October 2001. ISSN 0169-7439. .
15. Svein Jarle Horn, Einar Moen, and Kjetill Østgaard. Direct determination of alginate content in brown algae by near infra-red (NIR) spectroscopy. *Journal of Applied Phycology*, 11(1): 9–13, February 1999. ISSN 1573-5176. .
16. Anggara Mahardika, A. B. Susanto, Rini Pramesti, Hiroko Matsuyoshi, Bibin Bintang Andriana, Yusuke Matsuda, and Hidetoshi Sato. Application of imaging Raman spectroscopy to study the distribution of Kappa carrageenan in the seaweed Kappaphycus alvarezii. *Journal of Applied Phycology*, 31(2):1383–1390, April 2019. ISSN 1573-5176. .
17. Paul Geladi and Bruce R. Kowalski. Partial least-squares regression: a tutorial. *Analytica Chimica Acta*, 185:1–17, January 1986. ISSN 0003-2670. .
18. Enrique F. Schisterman, Albert Vexler, Brian W. Whitcomb, and Aiyi Liu. The Limitations due to Exposure Detection Limits for Regression Models. *American journal of epidemiology*, 163 (4):374–383, February 2006. ISSN 0002-9262. .
19. Tahir Mehmood, Kristian Hovde Liland, Lars Snipen, and Solve Sæbø. A review of variable selection methods in Partial Least Squares Regression. *Chemometrics and Intelligent Laboratory Systems*, 118:62–69, August 2012. ISSN 0169-7439. .
20. H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, December 1974. ISSN 1558-2523. . Conference Name: IEEE Transactions on Automatic Control.
21. Federico Marini and Beata Walczak. Particle swarm optimization (PSO). A tutorial. *Chemometrics and Intelligent Laboratory Systems*, 149:153–165, December 2015. ISSN 0169-7439. .
22. Y. Gong, J. Li, Y. Zhou, Y. Li, H. S. Chung, Y. Shi, and J. Zhang. Genetic Learning Particle Swarm Optimization. *IEEE Transactions on Cybernetics*, 46(10):2277–2290, October 2016. ISSN 2168-2267. .
23. Alankar Shrivastava and Vipin B Gupta. Methods for the determination of limit of detection and limit of quantitation of the analytical methods. *Chronicles of young scientists*, 2(1):21, 2011. ISSN 2229-5186. Publisher: Medknow Publications.
24. Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(Oct):2825–2830, 2011. ISSN ISSN 1533-7928.