# Semester project : Spectra analysis of PM2.5 for chemical speciation in monitoring networks

Author : Louise Aubet
Instructor : Satoshi Takahama
January 12, 2021 EPFL, Switzerland

**Abstract**

The Fourier transform infrared (FT-IR) spectra of fine particulate matter ($PM_{2.5}$) can be used to determine the corresponding concentration of different compounds. This project focuses here on ammonium sulfate. In this work, the strategy is to build a calibration model using both laboratory standards and collocated ambient measurements to build a common basis set. This method is a part of the transductive learning techniques and can be implemented using Principal Component Analysis (PCA) and Lasso regression. This model is compared to a base case with a Partial Least Squares (PLS) model trained on laboratory samples only, and two PCA and PLS ideal models that use ambient and laboratory samples for training. Upsampling techniques are needed to upsample the laboratory subset in order to have as many laboratory samples as ambient samples. Results show that the transductive model performances are better than the base case but they are not good enough for the model to be used operationally. However, the model succeeds in isolating the principal components corresponding to ammonium sulfate. Also, comparing two different PCA models suggests that higher performance could be reached with less sparse Lasso vector.

## 1 Introduction

Particulate matters or PM are fine particles of solid or liquid matter, suspended in the air. This is a category of particles that regroup many different chemical species, of different sizes, masses, etc. $PM_{2.5}$ refers to atmospheric particulate matter that have a diameter of less than 2.5 $\mu$m. They are the most dangerous as they penetrate further in the respiratory system and therefore can induce cardiovascular and pulmonary disorders as well as contribute to other adverse health effects, as it has been shown by S. Feng et al. [1]. $PM_{2.5}$ has started to become a public health issue and accurate measurements become central to better understand emissions, processes and impacts.

### 1.1 Prior art in chemical speciation

As PM display a large variety of chemical compounds, no instrument alone can distinguish the different species and measure their properties. For example, in the USA, the different monitoring networks use 3 filter types and 5 analytical methods for PM characterisation. The PTFE filters are used for gravimetry to measure the mass, for X-ray fluorescence to distinguish the different elements and for hybrid integrating plate and sphere to measure light absorption. Nylon filters are analysed using ion chromatography for inorganic ions detection. Quartz filters are analysed using thermal optical reflectance (TOR) for the presence of organic and elemental carbon [2].

## 1.2 IR Fourier transform and chemometrics

The use of Fourier transform can simplify this analysis. Vibrational modes of molecules happen mostly in infra-red so the major constituents that comprise $PM_{2.5}$ are represented in the infrared spectrum. Therefore, applying an IR Fourier transform to PTFE filters leads to spectra in which absorption intensity is related to bond abundance. One benefit of this technique is that PTFE samples are analysed non-destructively. The Fourier transform infrared (FT-IR) spectra of fine particulate matter contain many important absorption bands relevant for characterising the different compounds. PM characterisation can be done by analysing both FT-IR spectra and concentrations. These measurements allow to support understanding of sources and health impacts as well as other research goals. The idea is to build calibration models that learn how to interpret FT-IR spectra and the corresponding concentration and then are able to predict the concentration for new spectra. One way to do so is to use mixtures generated in laboratory for which we know both the spectrum and the concentration. The model trains on these samples and then predict by extrapolation the concentration of ambient samples using their spectra. The problem of this extrapolation is that it can generate shifts or broadening of absorption peaks. A more recent method is to use collocated field measurements. The concentration is measured using thermal optical reflectance (TOR) on PTFE filters. This second method has many benefits among which better results as the training samples are more complex [3]. As an example, ambient samples can be composed of 1000s of compounds whereas laboratory standards are composed of 1 to 3 compounds at most. One disadvantage of this method is that it implies to have TOR measurements on a large scale to get a sufficiently large training set. The same problem would arise when recalibration new samples.

## 1.3 Transductive learning

A solution to solve the drawbacks mentioned above would be to use ambient samples for the selection of the model features only and to combine them with laboratory samples. By doing so, the size of the necessary ambient data set would be smaller. The aim of this project is to build a calibration model that would use both laboratory and ambient samples. By comparing its performances with the ones of a base case model, one could study the advantages and disadvantages of such a model. This idea is part of multitasks or transductive learning methods in the sense that it uses specific cases for training in order to predict on other specific cases. As shown by O. Chapelle et al. [4], it aims to avoid the need of obtaining new samples for recalibration by searching a common feature representation. Using transductive learning may also make the calibration model more robust in extrapolation. For the sake of simplicity, we focus in this work on prediction of ammonium sulfate only. This choice is justified by the fact that ammonium sulfate is a major component of $PM_{2.5}$ [5]. It is also the most common form of sulfur containing species, and sulfur is an atmospheric aerosol that contributes to environmental and health related problems. The main research question is : Can a better calibration model be built from a set composed of both laboratory and ambient samples ? To be more precise : Does this new model extrapolate better than the base case model ? In other words, can the accuracy be improved ? Is having information about the ambient samples in the basis set of the calibration samples informative for model selection ? That is, does cross validation give better indication of actual performance in this case ?

Transductive learning includes both laboratory and ambient samples to derive the basis set common to both types of samples. However, this composite set is imbalanced as we have more ambient samples than laboratory samples. This could potentially negatively impact the calibration model. As ambient samples are more numerous, the model might bias towards the majority class and neglect the minority class. One method to improve performance is to balance out the number of examples between the two classes. There are two possible solutions : downsampling the ambient set or upsampling the laboratory set. The downsampling procedure on these particular data sets has already been previously discussed by Charlotte Bürki [6] and by C. Bürki et al. [7]. Also, removing samples from the majority class brings the risk of losing any useful data for the learning process. Therefore, we focus here on the upsampling procedure.

### 1.4 Upsampling technique

Resampling data sets is a technique that is sometimes used in the analysis of PM samples. This type of data has usually a skewed distribution, i.e. a high number of low-concentration samples and only a few high-concentration samples. However, high-concentration samples are very important as these are the ones that will have the biggest impact on health and climate. These are the concentrations that are the most likely to exceed the regulatory limits. One strategy to resample the data is to use Synthetic Minority Over-sampling Technique (SMOTE), proposed by N. V. Chawla et al. [8]. It oversamples by creating new data points, mid-way between two near neighbours in the minority class. In a sense, SMOTE fills in the feature space between points in the minority class. This technique is often used to solve this imbalance in classification problems but not only. For example, it is used by B. Gong and J. Ordieres-Meré [9] to predict daily maximum ozone threshold exceedances. A similar method is used by S. Park et al. [10], [11] but without using SMOTE algorithm explicitly. In this study, ground-level PM concentrations are extrapolated to estimate surface concentration over South Korea. They assumed that the concentration in one pixel should be close to the one of the neighbouring pixels.

Tough, this extrapolation using SMOTE is hard to interpret because it would create artificial spectra with peaks that didn't exist before. Therefore, this technique is not interpretable and robust enough to be used as a reference method. It is preferable to use naive upsampling. The idea is to supplement the data with multiple copies of some of the minority class, as done by C.-M. Vong et al. [12], such that both classes have the same number of samples. This is a simpler method but one question remains. How to choose the samples to replicate ? They can be chosen randomly, using venetian blinds or following another probability distribution. An example would be to replicate with a higher probability the most certain data points. Unfortunately, no studies were found concerning how to oversample by taking into account the uncertainty of the data points. Also, in the data set used in this project, there is no uncertainty measures so hypotheses would have to be made. The probability of choosing one data point can be inversely related to the variance of this point. A reasonable hypothesis would be to suppose that the variance increases with the absolute value of the concentration. A threshold at a minimum variance value would be set so very small values. This hypothetical method would automatically lead to the creation of more low concentration samples, that are already the more numerous.

## 2 Methods

In this section, we describe the model building process following the pathway described by S. Takahama et al. [3]. The first step is the construction of several models using the basis set. The second step is the evaluation of the models. All models with different parameters are evaluated using the same criterion. The best model is selected as the one having the best score with respect to the chosen criterion. The third step is the calibration : we train the model on the training data set. The final step is the prediction, when we use the model to predict concentration for the test set.

## 2.1 Model building

The aim of the calibration model is to predict concentrations of chemical element in an atmospheric mixture, after a training phase. For this type of task, linear predictive models are typically used, as they are robust and interpretable. Mathematically, let's $\mathbf{X}$ be the matrix containing the spectral measurements and $\mathbf{y}$ the vector containing the concentrations :

$$\mathbf{X} \in \mathbb{R}^{n \times m}, \mathbf{y} \in \mathbb{R}^{n} \tag{1}$$

where $n$ is the number of samples and $m = 2784$ is the number of variables, i.e. the number of wavenumbers. The goal is to find a solution to the following linear model :

$$\mathbf{y} = \mathbf{X} \cdot \mathbf{b} + \mathbf{e} \tag{2}$$

where $\mathbf{X}$ and $\mathbf{y}$ are given, $\mathbf{b}$ is the vector of linear coefficients and $\mathbf{e}$ is the error. The data in $\mathbf{X}$ is collinear, i.e. absorbances corresponding to close wavenumbers are dependent. Also, it is high-dimensional, i.e. there are usually more variables than measurements. For these two reasons, one cannot use the classic least squares model for example. Today, one model is mainly used for this task : Partial Least Squares (PLS).

### 2.1.1 Base case : Partial Least Squares (PLS)

The Partial Least Squares model is useful to find linear relations between two multivariate data sets represented by matrices. It is particularly suited when the matrix has more variables than samples and when there is multicollinearity among values. By contrast, standard regression will fail in this type of problems. In this case, we want to find the relationship between the IR spectra and the concentrations. Each spectrum includes 2784 variables that are often correlated with each other. The implementation is done in Python, using `scikit learn` library and more particularly the function `PLSRegression` [13]. The idea is to project the spectra onto a space of smaller dimension. Both $\mathbf{X}$ and $\mathbf{y}$ are projected onto basis sets $\mathbf{P}$ and $\mathbf{q}$.

$$\begin{cases} \mathbf{X} = \mathbf{T} \cdot \mathbf{P}^{T} + \mathbf{E}_{X} \\ \mathbf{y} = \mathbf{T} \cdot \mathbf{q}^{T} + \mathbf{e}_{y} \end{cases} \tag{3}$$

$\mathbf{P} \in \mathbb{R}^{m \times p}$ and $\mathbf{q} \in \mathbb{R}^{p}$ are usually called the loadings, $\mathbf{T} \in \mathbb{R}^{n \times p}$ is an orthogonal matrix called the score matrix, $\mathbf{E}_{\mathbf{X}} \in \mathbb{R}^{n \times m}$ and $\mathbf{e}_{\mathbf{y}} \in \mathbb{R}^{n}$ are the residuals. The decompositions of $\mathbf{X}$ and $\mathbf{y}$ are made so as to maximise the covariance of $\mathbf{T}$ with respect to $\mathbf{y}$ [14] [15]. It uses a latent variable approach as it reduces the number of variables to a smaller and uncorrelated set and performs least squares regression on these components. Practically, `PLSRegression` uses NIPALS algorithm or Nonlinear Iterative Partial Least Squares. The PLS model has one parameter that needs to be set by the user : the number of components $p$, i.e. the number of variables kept during the change of basis. The data is not scaled to preserve the original distribution of the samples.

To obtain the regression coefficient $\mathbf{b}$, the following equation is used :

$$\mathbf{b} = (\mathbf{P}\mathbf{P}^{T})^{-1}\mathbf{P}\mathbf{q}^{T} \tag{4}$$

with the loading matrices from the PLS decomposition.

### 2.1.2 Transductive learning : Principal Component Analysis (PCA)

The transductive learning method is better implemented using Principal Components Analysis (PCA) which includes both laboratory and ambient samples to derive the basis set common to both types of samples. PCA is another option for working with multivariate and collinear data. As its name suggests, PCA computes the principal components of $\mathbf{X}$ which is a sequence of vectors $\{\mathbf{w}_k\}_{k=1}^{m} = \mathbf{W}$, where $w_i$ is the direction of the line that best fits the data and that is orthogonal to the first $i-1$ vectors. The set $\{\mathbf{w}_k\}_{k=1}^{m}$ forms an orthonormal basis in which the dimensions of the data are now linearly uncorrelated. A PCA model computes the principal components and then perform a change of basis on the data. Often only the first few principal components are used, especially when the number of dimensions is high. This gives a truncated transformation. Contrary to PLS, PCA only performs a change of basis and doesn't apply a linear regression. Therefore, a Lasso regression is applied after the principal component analysis. Lasso regression (or Least Absolute Shrinkage and Selection Operator) combines both mean squared error (MSE) cost function and $L_1$-regularization. It is defined by the following formula :

$$L(\mathbf{q}) = \min_{\mathbf{q}} \frac{1}{n} \sum_{i=1}^{n} \left(\mathbf{y_i} - \mathbf{X}_i^T \mathbf{q}\right)^2 + \alpha ||\mathbf{q}||_1 \tag{5}$$

where $||\mathbf{q}||_1 = \sum_{i=1}^{n} |q_i|$, $\alpha$ is the regularisation parameter set by the user and $\mathbf{q}$ is the vector of coefficients of the linear model. This regression method performs both variable selection and regularisation at the same time. Therefore, the variable selection, i.e. dimension reduction, can be done either by truncation of the matrix or by applying Lasso. In this work, we keep all the components in PCA and then use Lasso regression to select the variables. The smallest eigenvalues correspond to error and instrument noise, so by dropping them Lasso makes the model less sensible to noise. This ability of Lasso to exclude useless variables from equations makes it a little bit better than Ridge regression at reducing variance in models that contain a lot of potentially useless variables.

To obtain the regression coefficient $\mathbf{b}$ on the original variables, Equation (4) from PLS can be used again. The loading vector $\mathbf{q}$ is replaced by the sparse vector $\mathbf{q}$ containing the coefficients of the Lasso regression. The loading matrix $\mathbf{P}$ is obtained by multiplying the columns of $\mathbf{W}$ by the square root of the corresponding eigenvalues, i.e. the corresponding variances.

### 2.1.3 Intermediate models

The two models, PLS for base case and PCA for transductive learning, cannot be compared directly as they differ in both the type of model and the data used. Therefore, let's define two intermediate models as summarised in Table 1. Model M2 can be compared to M1 as they differ only by the data used. M3 can be compared with M2 as they differ by the model used. And finally, M4 can be compared to M3 as they differ by the data sets used. M2 and M3 are not realistic models. They describe an idealistic case if we had the necessary ambient samples for the training phase. This is the best case scenario and we expect that M4 score will be between M1 score and M2/M3 score.

Table 1: Experimental design of $PM_{2.5}$ characterisation using transductive learning on FT-IR spectra.

| | | Model features / Basis set | |
|---|---|---|---|
| | | Laboratory only<br>(Base case) | Laboratory + Ambient<br>(Transductive case) |
| Training | Laboratory only | M1 : PLS | **M4 : PCA + Lasso (objective)** |
| data set | Laboratory + Ambient | - | M2 : PLS<br>M3 : PCA + Lasso |

## 2.2 Model selection

As it has been seen, each model has one parameter. The best model has to be selected by optimising the value of the parameter. This optimisation problem can be solved by computing the performance of the models with different values of the parameter and choosing the one that maximises the performance. Also, we would like to select a robust model, i.e. not a model that gives accurate predictions on the known training data but rather one that gives good predictions on the new data and avoids overfitting or underfitting. In general, we can't know how well a model will perform on new data until we actually test it. To address both these problems, we can perform a cross-validation. Therefore, cross-validation serves two main functions : assessing the optimal complexity of the model and estimating the performance of the model applied to unknown data. The idea is to split several times the initial training data set into separate calibration and test subsets. Then permutations are performed such that the model is trained and validated on different subsets over a repeated number of trials. This leads to a distribution of score metrics that are then averaged to obtain the final score of the cross-validation. The metric used is the root mean squared error (RMSE) between the prediction of the model and the reference value. For each trial, the RMSE value is computed and then they are averaged over the number of trials. Furthermore, we want to develop a strategy that we can use when particular ambient reference samples are not available. Therefore, in model M4, the cross-validation is run on a common set composed of ambient and laboratory samples, to determine the optimal model features. The training set is composed of laboratory samples only though.

The separation between calibration and validation subsets in the cross-validation can be done in different ways. In this work, we are not using a random sampling but rather another method called Venetian blinds. It can be also referred to as interleaved or striped cross-validation. With this method, each subset is determined by selecting every $n$-th object in the data set. If the total number of samples is $n$ and the number of splits is $s$ then there will be $s$ subsets of $n/s$ samples each. The number of splits or folds is a parameter of the cross-validation. Figure 1 provides a visual description of this process. The rest of the cross-validation is similar to the classic $k$-fold : among the obtained subsets, keep one as validation set and the others form the calibration set. This method is better than randomisation when the distribution is log-normal because it avoids the problem of having a lot of outliers in the same subset. It is also reproducible.

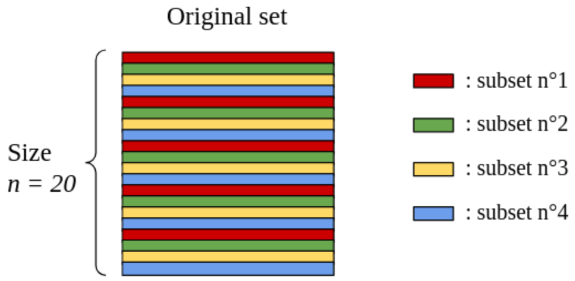

Figure 1: Scheme of an example of venetian blinds splitting with $n = 20$ and $s = 4$.

## 2.3 Model evaluation

Each model is evaluated using its performance on the test set. A typical graph displays the prediction value on the $y$-axis with respect to the reference value on the $x$-axis. The prediction should then be as close as possible to the reference value, i.e. it should lie on the $y = x$ line if the model was perfect. Therefore, the performance of the model can already be estimated at a glance. To be more precise, different metrics can be used. Following the suggestions of S. Takahama et al. [3], we are using in this work the root mean square error (RMSE) defined as :

$$RMSE(\hat{\mathbf{y}}, \mathbf{y}) = \sqrt{< (\hat{\mathbf{y}} - \mathbf{y})^2 >} \tag{6}$$

as well as the mean and median bias defined as :

$$\text{Mean bias}(\hat{\mathbf{y}}, \mathbf{y}) =< \hat{\mathbf{y}} - \mathbf{y} >, \qquad \text{Median bias}(\hat{\mathbf{y}}, \mathbf{y}) = \text{Med}[\hat{\mathbf{y}} - \mathbf{y}] \tag{7}$$

where $\mathbf{y}$ is the vector of reference values, $\hat{\mathbf{y}}$ is the vector of prediction values, $< \cdot >$ is the sample mean and Med[$\cdot$] is the sample median. Median bias is a robust metric whereas mean bias is sensible to extreme values that may lead to misinterpretation.

# 3 Experimental data

## 3.1 Description of the data

The data consists of 4 data sets: {ambient, laboratory} × {spectra, reference measurements}. They are rds, csv and txt files so we need to convert them all into csv files to process them using Python. The measurements for the ambient data set come from the monitoring network IMPROVE in the US and have been collected during the period $2011 - 2013$.

A spectrum is the measure of the absorbance $A$ [unitless] as a function of the wavenumber $w_n$ [cm$^{-1}$]. Here, the wavenumber goes from approximately 420 to 3998 cm$^{-1}$, for a total of 2784 measures. This corresponds to radiation from 2500 to 25000nm. The data points are evenly spaced in inverse meters and there is no missing value. They come from the FT-IR spectroscopy of PM samples collected on PTFE filters. Spectra are already baseline-corrected by subtracting the background spectra and scattering contributions from the PTFE filter. So the spectral data is in the form of a matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$ where $n$ is the number of samples and $m = 2784$ is the number of wavenumbers. The concentrations of various chemical elements are given in [$\mu$g/cm$^2$]. They come from TOR reference measurements. For ambient samples, the concentrations have been measured at the same position as the spectrum, i.e. they are collocated measurements. The ambient reference measurements are displayed in a data table containing for each samples, the concentration in [$\mu$g/cm$^2$] of various compounds. Laboratory-standard measurements are also displayed in a data table. For each sample, the type of compound, the type of data set (train or test) and concentrations for various functional groups are given in [$\mu$mol/cm$^2$]. There are in total nine types of organic compounds and there are also organic blanks.

## 3.2 Cleaning of the data

As explained in section 1.3, this work focuses on ammonium sulfate ($(NH_4)_2SO_4$) so we want this concentration in particular. However, the concentration of ($(NH_4)_2SO_4$) is not given as such. Therefore, we use the concentrations of functional groups from which we can approximate the concentration of ammonium sulfate. Ambient measurements contain concentrations of ion sulfate $SO_4$, in [$\mu$g/cm$^2$]. So we need to apply the following transformation using molecular masses :

$$C_{(NH_4)_2SO_4} = C_{SO_4} \cdot \frac{M_{(NH_4)_2SO_4}}{M_{SO_4}} \tag{8}$$

with $M_{(NH_4)_2SO_4}$ [g/mol] and $M_{SO_4}$ [g/mol] the molecular mass of ammonium sulfate and ion sulfate respectively. By doing so we assume that all sulfate is in the form of ammonium sulfate which is a good hypothesis. Laboratory standards contain the concentration of NH, in [$\mu$mol/cm$^2$]. We can deduce the corresponding concentration of ammonium sulfate using the equation :

$$C_{(\mathrm{NH_4})_2\mathrm{SO_4}} = C_{\mathrm{NH}} \cdot M_{(\mathrm{NH_4})_2\mathrm{SO_4}} \cdot \frac{n_{(\mathrm{NH_4})_2\mathrm{SO_4}}}{n_{\mathrm{NH}}} \tag{9}$$

where $n_{(\mathrm{NH_4})_2\mathrm{SO_4}}/n_{\mathrm{NH}} = 1/8$ is the number of moles of NH bond per mole of $(\mathrm{NH_4})_2\mathrm{SO_4}$.

If some samples contain NaN values in the measurements, they are removed. Moreover, only samples for which both spectrum and concentration are available, i.e. collocated measurements, are kept. A sample corresponds to a particular site and date. The list of sample labels is extracted from both the spectra and the concentrations. These two lists of labels are compared and only the intersection is kept, i.e. the samples for which both spectrum and concentration were measured on the same site and the same date.

Figure 2 displays the distributions of ambient and laboratory samples after data processing. For both sets the distributions are right-skewed, that is they are not symmetric and they have a long right tail. The range of the ambient samples was broader as maximum concentration was around 300 $\mu$g/cm$^2$ whereas for laboratory samples it was around 130 $\mu$g/cm$^2$. This discrepancy might affect the performance of the model as it would need to extrapolate from smaller laboratory concentrations to higher ambient concentrations. In this project, we don't want to study the ability of the model to extrapolate but rather the impact of transductive learning method. This could be a potential problem as we wouldn't know if a particular phenomenon is due to the extrapolation or to the model. Therefore, we decide to exclude ambient samples for which concentration is superior to 130 $\mu$g/cm$^2$. Therefore, it can be seen in Figure 2 that both distributions span the same interval of concentrations. The data is not standardise because this would modify the concentration distribution of the samples. Keeping the original density function is important when training the model and interpreting the results. After data cleaning, the two sets have the following size : 2214 ambient samples and 218 laboratory samples.

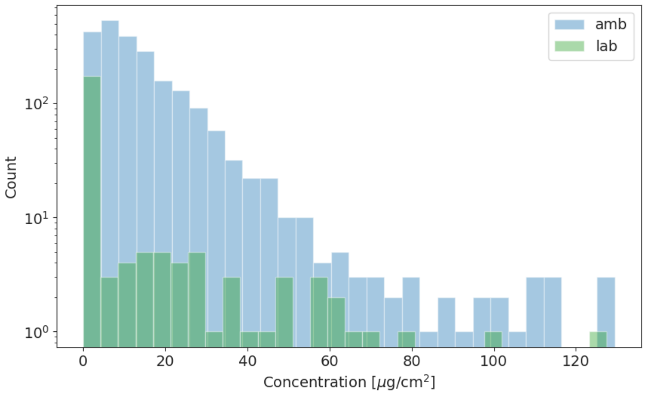

Figure 2: Histogram of the distributions of ambient and laboratory samples after data processing.

# 4 Results

## 4.1 Upsampling

Upsampling techniques have been discussed in section 1.4. In this work, upsampling is necessary as the laboratory data set contains 218 samples whereas the ambient data set contains 2214 samples. Therefore, the laboratory data set must be upsampled to reach the same size as the ambient data set. It must be upsampled by a factor 10 approximately. We choose here to use a naive upsampling technique. The simplest solution would be to randomly choose the data points to oversample. Taking into account the uncertainty of the data points could lead to a more balanced resampled data set, but there is no uncertainty measures in the used data set, so hypotheses would have to be made. The probability of choosing one data point can be inversely related to the variance of this point. A reasonable hypothesis would be to suppose that the variance increases with the absolute value of the concentration. This hypothetical method would automatically lead to the creation of more low concentration samples, that are already more numerous. This consequence would increase the ratio of low concentration samples over high concentration samples. Therefore, for this work, the data points to oversample are chosen randomly. However, this method could give resulting sets with different distributions than the original ones. Therefore, let's first verify that the distributions of the upsampled sets are similar to the corresponding original sets. In order to do so, we plot the histograms of the distribution of ammonium sulphate in Figures 3 and 4.

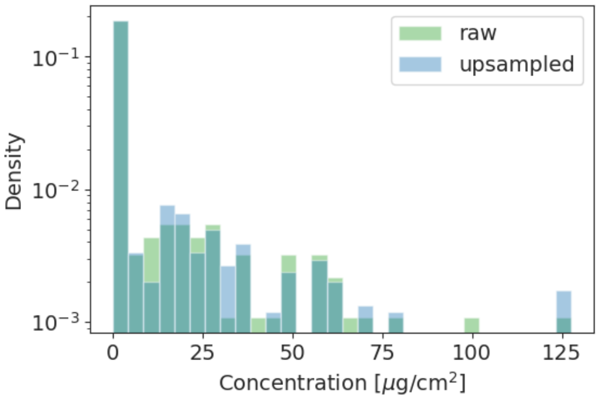

Figure 3: Histogram of the distributions of laboratory samples before and after upsampling.

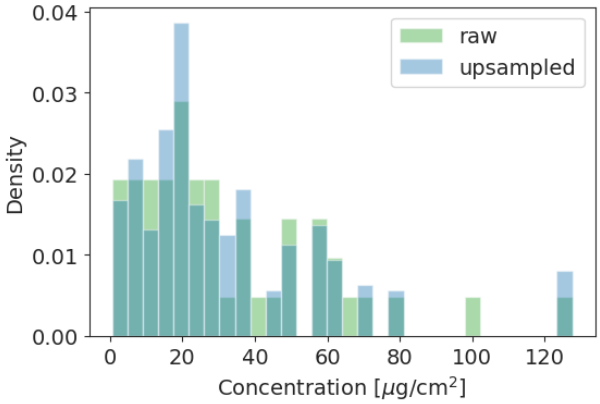

Figure 4: Histogram of the distributions of non-zero laboratory samples before and after upsampling.

In Figure 3, a peak for zero concentrations can be observed, and it tends to flatten the rest of the histogram, that's why Figure 4 is plotted with non-zero concentrations only. In both figures, the distributions are very similar as it can be seen that the histograms almost perfectly overlap each other. The same analysis is applied to the concentration of aCH, that was also indicated in the files containing informations about the laboratory samples. The same results are obtained : the distributions before and after upsampling are very similar. It seems that randomly resampling by a factor of 10 doesn't modify the distributions of the data.

## 4.2 Comparison in prediction

Table 1 presents the four models that are evaluated in this work and that have been described in section 2. Let's start by comparing the predictions of each model with each other. All four models are selected among models with parameter $p \in [2, 50]$, $p \in \mathbb{N}$ for PLS and $\alpha \in [10^{-9}, 10^{-2}]$, $\alpha \in \mathbb{R}$ for PCA. The best model is selected using venetian blinds cross-validation with 5 folds. Figures 5 to 8 display the prediction values as a function of the reference values. RMSE errors and biases are separated for ambient and laboratory samples and are summarised in Table 2. The error that is the most interesting for future applications is the ambient error as the final objective is to predict concentration for new ambient samples. Graphically, we can clearly see that models 2 and 3 are very good and Table 2 confirms that fact since the error of M2 and M3 is approximately 5 times smaller than the error of the base case and 2.7 times smaller than the error of M4. The results concerning biases are similar. The mean bias for M4 is approximately 4 times smaller than the one for M1 but 10 times bigger than the ones for M2 and M3. These extremely good results for models M2 and M3 were expected as they are ideal cases in which the model is trained with ambient samples. The biggest RMSE for ambient samples is reached for model M1. This is also expected as this model hasn't seen any ambient samples during the training or basis set construction phase. The ambient samples are very different from laboratory samples because they are more complex and the model can't handle well this increase in complexity. The RMSE for ambient samples in model M4 is only 1.8 times smaller than the RMSE for the base case, which is not a remarkable improvement. Graphically, it can be seen that there a lot of points far from the diagonal, which means bad predictions.

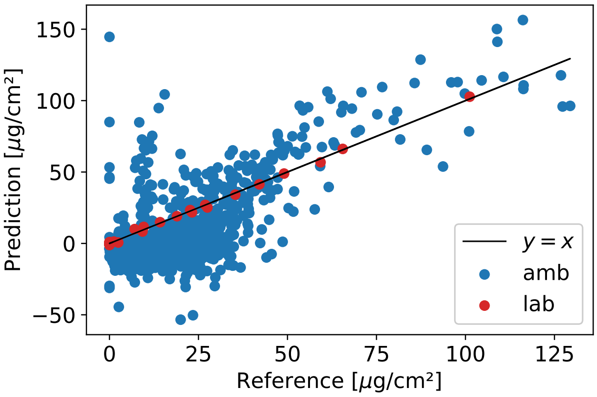

Figure 5: Prediction of M1 model as a function of reference measurements.

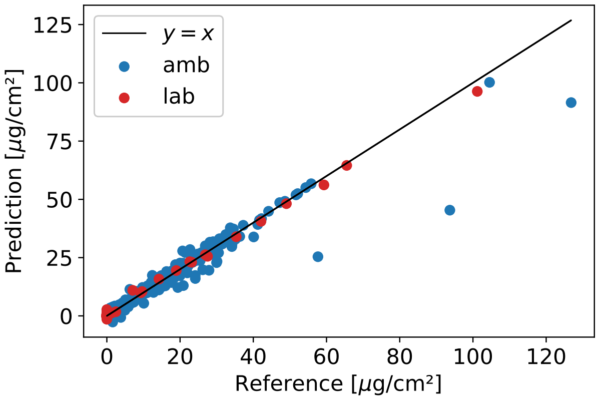

Figure 6: Prediction of M2 model as a function of reference measurements.

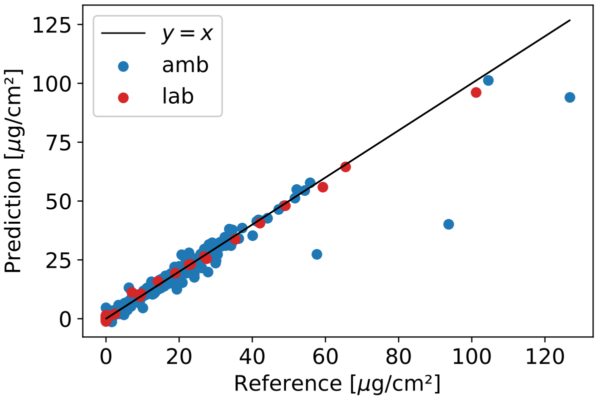

Figure 7: Prediction of M3 model as a function of reference measurements.

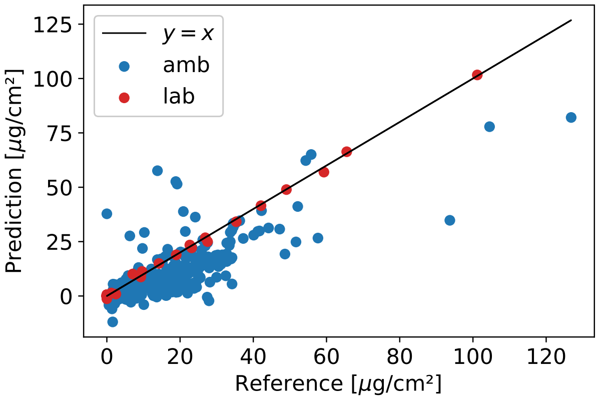

Figure 8: Prediction of M4 model as a function of reference measurements.

Table 2: Table summarising RMSE and biases of the four models for ambient and laboratory samples.

| | M1 | M2 | M3 | M4 |
|---|---|---|---|---|
| RMSE amb | 18.356 | 3.667 | 3.645 | 9.901 |
| RMSE lab | 0.768 | 1.061 | 1.048 | 0.706 |
| Mean bias amb | 11.895 | 0.323 | 0.323 | 3.378 |
| Mean bias lab | 0.0131 | 0.0749 | 0.0973 | 0.0721 |
| Median bias amb | 12.243 | −0.0451 | −0.0586 | 4.317 |
| Median bias lab | 0.00697 | 0.0479 | 0.03901 | 0.0188 |

Let's investigate further these outlier points in model M4. They mainly come from 2 different sites. Figure 9 displays the outlier points and their sites : BYIS1 and FRES1. They can also be seen in the graph for model M1. Furthermore, in models 2 and 3, there are three points that are still very badly predicted. They also come from the site BYIS1. BYIS1 is a site situated in South Korea, so the background is very different than the other samples that come from the US. Most of the highest concentrations come from this site as high concentrations of sulphur dioxide are transported from China, emitted by coal burning. These reasons explain why the model has some difficulties to predict the concentration for samples coming from BYIS1, even in the case of models 2 and 3 where the training phase includes ambient samples. The other site, FRES1 is situated in Fresno, in California. The absence of coal burning in this area implies low sulphur concentrations and therefore low ammonium sulfate concentrations. However, the samples are collected in a urban area whereas IMPROVE is mostly a rural monitoring network with only a few urban sites. Therefore, the presence of a lot of vehicles that are emitting organic compounds for example creates very particular spectra and this might explain the discrepancy for FRES1. More precisely, increasing the amount of organic compounds makes the "ammonium sulfate/organic matter" signal to noise ratio decrease. This decrease creates interferences and finally the model performance decreases.

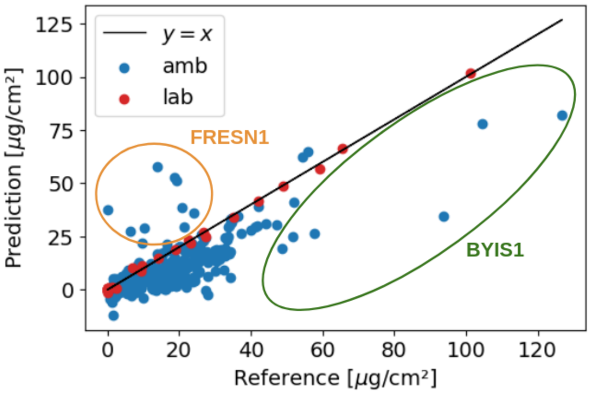

Figure 9: Prediction of M4 model as a function of reference measures, with sites.

### 4.3 Comparison of model selection

We have seen that model M4 performs better than model M1 as expected but its performances are still far from the performance of M3. The performances of M4 are not good enough for it to be used operationally as it is. Let's investigate further the model selection process in order to better understand this transductive model.

Let's compare the calibration and the model selection process. For each model with a different parameter, the RMSE of the cross-validation (RMSECV), the RMSE of the calibration (RMSEC) and the RMSE of the prediction (RMSEP) are computed. Figures 10 to 13 illustrate the graphs obtained with the evolution of the three errors with respect to the parameter of the model. Figures 10 and 11 describe the PLS models and Figures 12 and 13 describe the PCA models. The behaviour of the curves are similar within these pairs. In Figures 10 and 11, the RMSEC curves decrease as the number of latent variables in the model increases and is very smooth. This is intuitive because as the number of variables increases, the model trains on more data and therefore learns better. As explained in section 2.2, cross-validation allows us to estimate the performance of the model applied to unknown data. Therefore, the validation samples are not used to build the model that is used to test them. This explains why the RMSECV curve undergoes more oscillations, the curve is less smooth than RMSEC. Also, it is not monotonically decreasing, the value can increase as too many latent variables are added to the model. In the PLS models for example, increasing the number of components implies better predictions at first but at one point, overfitting can appear and make the RMSECV increase. The optimal number of latent variables is taken as the number for which the RMSECV is minimum. It is indicated by a red cross. In Figure 10 the minimum of RMSECV is reached for 29 components and in Figure 11 it is reached for 44 components. For model M1, the RMSEP is increasing with the number of latent variables and is one order of magnitude higher than RMSECV. For model M2, the RMSEP curve is closer to the RMSECV curve : they have both a similar behaviour, RMSEP is only approximately 1.3 times higher than RMSECV. This difference in the RMSEP curve is due to the fact that model M1 is trained on laboratory standards only whereas in model M2, the cross-validation and the training is done using the common basis set.

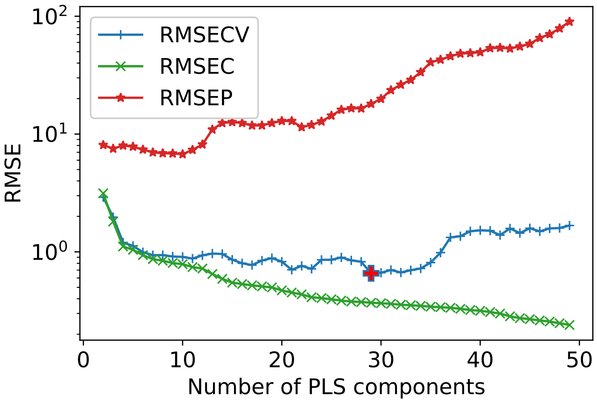

Figure 10: Graph of RMSECV, RMSEC and RMSEP for venetian blinds cross-validation on M1.

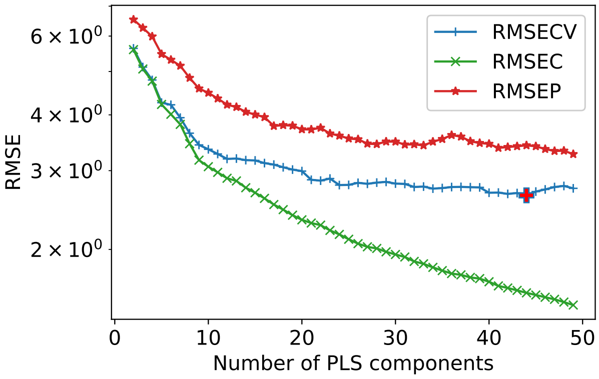

Figure 11: Graph of RMSECV, RMSEC and RMSEP for venetian blinds cross-validation on M2.

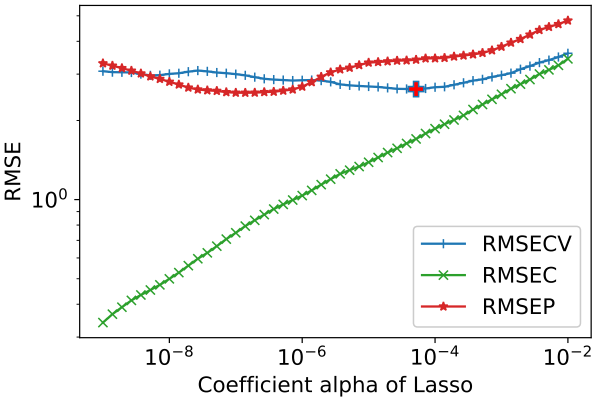

Figure 12: Graph of RMSECV, RMSEC and RMSEP for venetian blinds cross-validation on M3.

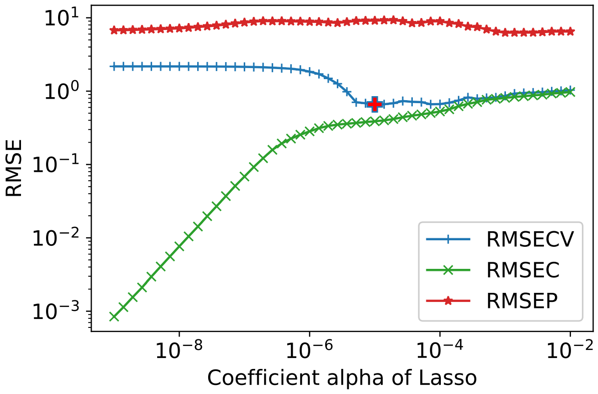

Figure 13: Graph of RMSECV, RMSEC and RMSEP for venetian blinds cross-validation on M4.

In Figures 12 and 13, the cross-validation is now done with respect to the regularisation parameter of Lasso. We can observe that in both cases the RMSEC is increasing when the coefficient $\alpha$ is increasing. The two RMSECV curves are decreasing and then increasing, leading to a minimum at $\alpha = 5.179 \cdot 10^{-5}$ for M3 and at $\alpha = 1.151 \cdot 10^{-5}$ for M4. In this case, the optimal number of components is very different so it seems that having ambient samples in the calibration sets is informative for model selection. The RMSEP curves have different behaviours as for the model M3, it increases when $\alpha$ is increasing, whereas for model M4 it decreases. Furthermore, for M4, the RMSEP is one order of magnitude higher than the RMSECV. A similar observation was done for M1 and might be due to the fact that these two models use only the laboratory samples during the training phase. During the prediction phase, the test samples are very different from the training samples as some of them are ambient samples. Therefore the RMSECV is overly optimistic compared to the RMSEP.

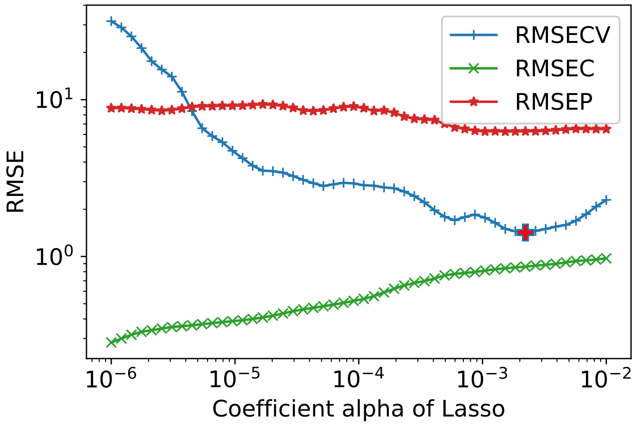

Figure 14: Graph of RMSECV, RMSEC and RMSEP for classic cross-validation on M4 with 5 folds.

Also, Figure 13 shows that the minimum of the RMSECV doesn't correspond to the minimum of the RMSEP. This puts into question the venetian blinds cross-validation as an appropriate criterion for this model selection. Let's repeat the experiment but using a classic cross-validation, separating randomly the samples. The error curves obtained are shown in Figure 14. The RMSECV curve differs from the one in Figure 13. Its minimum now gives an optimal parameter $\alpha = 2.223 \cdot 10^{-3}$. The resulting RMSEP is smaller than the one obtained with venetian blinds cross-validation, RMSE $_{\text{amb}}$ $= 6.778$, so the performance of the model is better. We can ask ourselves if this is just a lucky example or if the random cross-validation is really a better criterion that venetian blind cross-validation in this case. Cross-validation is one criterion for model selection among many so there are other possibilities. To overcome this problem of criterion for model selection, let's study M3 and M4 in more details by comparing their RMSEP.

## 4.4 Further comparison of M3 and M4

We have seen that model selection could be problematic in the sense that different criteria could lead to various optimal values for the parameter. In order to avoid this problem, we can scale up and consider the system with all the different prediction errors for the different values of $\alpha$. The RMSEP cannot be used to choose a particular value of $\alpha$ but it can be used a posteriori to better understand the behaviour of the models. M4 is the final model we are interested in and we would like its performance to be as close to the ideal model M3 as possible. Therefore, we can ask the following question : Can model M4 ever theoretically reach the performance of model M3 ?

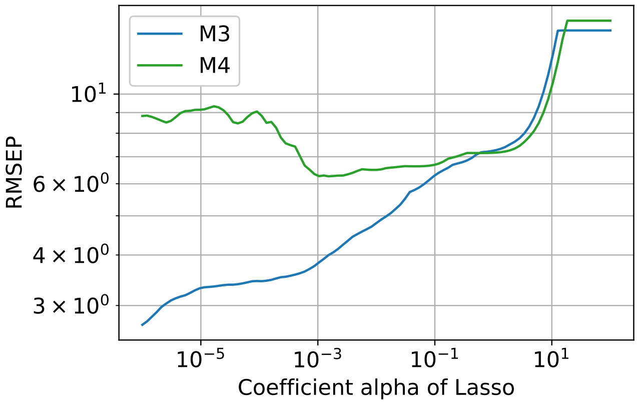

Figure 15: Comparison of RMSEP curve for model M3 and M4.

Let's study in more details the performance of model M3 and M4. Figure 15 illustrates the prediction error for both models, on the range of $\alpha$ values. This graph can be separated into three different parts. For $\alpha \geq 10$, there is a plateau, i.e. the RMSEP is constant and has similar values for both models. For $\alpha \in [1, 10]$, the curves almost overlap which means that for the same parameter, model M4 performs as good as M3. For $\alpha < 1$, the two curves start diverging. The prediction error for M3 continues decreasing while the prediction error for M4 plateaus then increases again. Let's have a look at the coefficients of the models. We can investigate the Lasso regression vector $\mathbf{q}$, the loading matrix $\mathbf{P}$ and the regression vector on the original variables $\mathbf{b}$. When $\alpha \geq 10$, all Lasso coefficients are zero : $\mathbf{q} = \mathbf{0}$. The error represents the averaged sum of squares of the concentrations. As $\alpha$ decreases, more and more Lasso coefficients become non-zero. For $\alpha \in [1, 10]$, the overall RMSEP of the two models are very close and the prediction points are very similar too. The Lasso vector is very sparse since only a few coefficients are non-zero, and are very similar between the two models. For $\alpha = 0.5$ for example, M3 keeps 3 Lasso coefficients and M4 keeps 2 coefficients. The values of the coefficients are $[17.403, -2.282, -1.793]$ for M3 and $[18.059, -2.599]$ for M4. For $\alpha = 10^{-2}$, M3 keeps 22 Lasso coefficients whereas M4 keeps 8 coefficients. The first two coefficients are similar between the two models and compared to the case $\alpha = 0.5$, but the rest of the coefficients differs. For $\alpha = 10^{-5}$, M3 keeps 156 Lasso coefficients whereas M4 keeps only 55 coefficients. The first two coefficients are very far from 18 and $-2$ as in the previous cases and the coefficients are very different between M3 and M4.

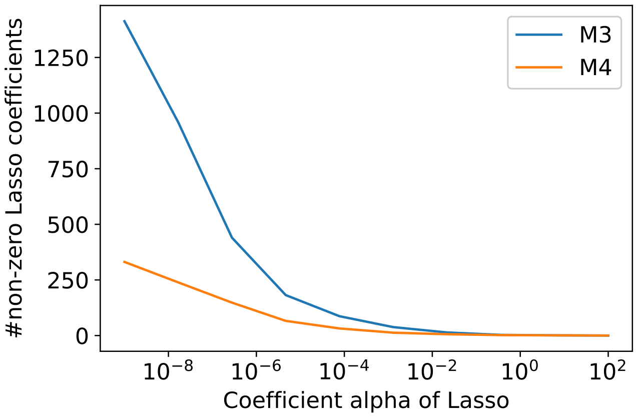

Figure 16: Number of non-zero Lasso coefficients as a function of $\alpha$.

Figure 16 displays the number of non-zero Lasso coefficients as a function of the value of the Lasso parameter $\alpha$. As $\alpha$ is decreasing, the number of non-zero coefficients is increasing, as it was observed above for some special cases. The regression models become more complex when $\alpha$ becomes small. The sparsity of vector $\mathbf{q}$ comes from the properties of Lasso : the 1-norm reduces the value of the coefficients and can set some of them to zero. Another interesting observation is that M3 keeps more coefficients than M4. As it has been shown before with some special cases the difference in the number of non-zero coefficients is very small between M3 and M4 for high values of $\alpha$ and is diverging for smaller values of $\alpha$. As M3 performs better and keeps more coefficients than M4 we can deduce that a less sparse vector $\mathbf{q}$ makes the model more robust during extrapolation. M4 discards too many features so the model is too simple and cannot extrapolate well during the test phase. This result suggests to reduce sparsity in model M4 in order to improve it. One idea could be to apply a regression that doesn't reduce the number of variables in the model, such as Ordinary Least Squares or Ridge regression. By using one of these models, we would keep all the coefficients and this would therefore lead to a model with a very high variance. Using the Ridge regression with a regularisation parameter should limit the problem of overfitting. As we have seen before, M3 kept more coefficients and also the values of the coefficients were not the same. Therefore, keeping all the coefficients in the regression might not be sufficient to reach a good performance. Evaluating the value for each coefficient should also be an important part of the process.

Let's now have a look at the loading matrix **P**. Models M3 and M4 have the same basis set, PCA is applied on a set containing both ambient and laboratory samples, and so they have the same loadings. Studying the loadings can bring an insight on the most useful wavenumbers.

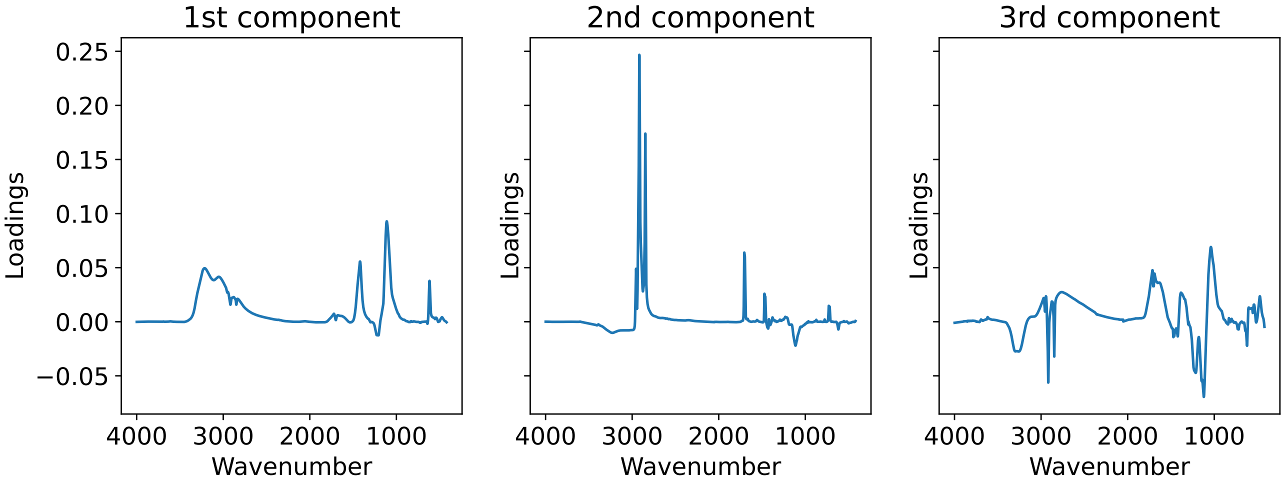

Figure 17: First three principal components from the loading matrix.

Table 3: Table summarising the 5 principal components that are the closest to the ammonium sulfate spectrum.

| **Rank** | **Component** | **Distance** |
|---|---|---|
| 1 | 1 | 63.7 |
| 2 | 3 | 70.1 |
| 3 | 5 | 71.4 |
| 4 | 6 | 72.4 |
| 5 | 50 | 72.8 |

Figure 17 shows the first three principal components for M4, corresponding to the first three rows of the loading matrix. In theory, the different principal components in PCA should isolate different chemical species. The peaks that correspond to the same species are covariant so they should be attributed to the same component and not be separated between different components. In this case, the first principal component looks like the ammonium sulfate spectrum [16]. A broad and prominent peak can be observed around 3000 $\text{cm}^{-1}$ and corresponds to the NH stretch. Two sharp peaks are situated around 1000 and 1500 $\text{cm}^{-1}$ and could correspond to the S=O and NH stretches respectively. The second principal component looks like a hydrocarbon spectrum. We can observe a sharp peak around 3000 $\text{cm}^{-1}$ that corresponds to the –$CH_3$ stretch. A way to evaluate the similar between the spectra is to compute the euclidean distance between the different principal components and the ammonium sulfate spectrum. The principal components can be ranked in order of decreasing similarity. The spectra are first normalised and then the 2-norm of the difference is computed. Table 3 summarises the results of the process. The 5 most similar principal components are displayed with the corresponding rank and the distance. As expected, the first principal component is very similar to ammonium sulfate, it is actually the closest component. Then, the third component, also displayed in Figure 17 comes in second position. However the similarity with the ammonium sulfate spectrum is already less obvious. The peaks are weaker and there is more noise. This is reflected in Table 3 as the distance between the spectra has increased by 10% compared to the first component. Therefore, it seems that the model succeeds in isolating the ammonium sulfate characteristics. It is the most important compound in this case as the model focuses on predicting the concentration of ammonium sulfate. Also the first principal component has the highest variance and therefore represents the main piece of information. This must be due to the laboratory standards that composed the training set. Indeed, the laboratory spectra only contain 2 or 3 chemical compounds and are less noisy than the ambient samples, and therefore it is easier for the model to isolate them.

# 5 Conclusion

We presented a new framework to predict concentration of ammonium sulfate in $PM_{2.5}$ samples, using the transductive learning technique. The idea is to use a common basis set composed of both laboratory standards and ambient samples. This can be implemented using a model composed of PCA and Lasso regression. The aim is to test whether a better calibration model could be built from this common basis. In order to do so, this final model has been compared to a PLS base case model using laboratory samples only, and two intermediate models : PLS and PCA using both laboratory and ambient samples in the training set. The model performs better than the base case that uses only laboratory samples, but it is still far from the performances of an ideal model using ambient samples in the training set. We would like the performances to tend towards this ideal model. However, the final model doesn't perform well enough to be used in practice. A lot of points are still badly predicted, especially for sites that are very different from the majority of sites in the monitoring network. Concerning the model selection process, the optimal number of components selected by cross-validation is very different between the base case and the transductive case. Therefore, it seems that having ambient samples in basis set is informative for model selection in this case.

Even tough the current performance of the final model M4 is not high enough for an operational use, we can further study the model in order to find the outline of potential improvements. Studying the Lasso coefficients reveals that the ideal model M3 discards less coefficients than M4 and this could be one of the reasons that explain the discrepancy between the performances of the two models. The next step would be to verify this hypothesis and modify M4 by using Ordinary Least Squares or Ridge regression instead of Lasso regression. These regressions would not discard any coefficients and this could potentially lead to an enhancement of the performances. A downside of this technique is the risk of overfitting. Furthermore, model M4 succeeds in isolating the different chemical compounds in the principal components. The first principal component corresponds to ammonium sulfate.

This work enables many directions for future studies. First, the analysis of the PCA model could be carried out further. A deeper analysis could investigate if a higher prediction accuracy can be achieved with fewer samples or if the prediction error is stable over a wide range of solutions. Another extension could be to apply the model to other data sets, for other compounds like organic carbon (OC) or dust for example. In conclusion, this work shows that the transductive model cannot be used as it is for the moment, although the properties of the model are encouraging and there exist some lines of improvement.

# References

Sources verified on January 12, 2021:

[1] S. Feng et al., *The health effects of ambient PM2.5 and potential mechanisms*, Ecotoxicology and Environmental Safety (Vol. 128), 2016

[2] Bruno Debus et al., *Towards a single filter, single analytical method speciated PM monitoring network*, Conference abstract, AGU Fall Meeting, 2019, `https://agu.confex.com/agu/fm19/meetingapp.cgi/Paper/518363`

[3] S. Takahama et al., *Atmospheric particulate matter characterization by Fourier transform infrared spectroscopy: a review of statistical calibration strategies for carbonaceous aerosol quantification in US measurement networks*, Atmospheric Measurement Techniques, 2018

[4] O. Chapelle et al., *Semi-Supervised Learning*, The MIT Press, Cambridge, 2010

[5] C. W. Lewis and E. S. Macias, *Composition of size-fractionated aerosol in Charleston, West Virginia*, Atmospheric Environment, 1980

[6] Charlotte Bürki, *PC analysis of ambient PM infrared spectra*, EPFL, February 2020

[7] C. Bürki et al., *Analysis of functional groups in atmospheric aerosols by infrared spectroscopy: method development for probabilistic modeling of organic carbon and organic matter concentrations*, Atmospheric Measurement Techniques, 2020

[8] N. V. Chawla et al., *SMOTE: Synthetic Minority Over-sampling Technique*, Journal of Artificial Intelligence Research (Vol. 16), 2002

[9] B. Gong and J. Ordieres-Meré, *Prediction of daily maximum ozone threshold exceedances by preprocessing and ensemble artificial intelligence techniques : Case study of Hong Kong*, Environmental Modelling  Software, June 2016

[10] S. Park et al., *Estimation of ground-level particulate matter concentrations through the synergistic use of satellite observations and process-based models over South Korea*, Atmospheric Chemistry and Physics, January 2019

[11] S. Park et al., *Estimation of spatially continuous daytime particulate matter concentrations under all sky conditions through the synergistic use of satellite-based AOD and numerical models*, Science of the Total Environment, January 2020

[12] C.-M. Vong et al., *Predicting minority class for suspended particulate matters level by extreme learning machine*, Neurocomputing, 2012

[13] Pedregosa et al., *Scikit-learn: Machine Learning in Python*, JMLR 12, 2011, `https://scikit-learn.org/stable/modules/generated/sklearn.cross_decomposition.PLSRegression.html#sklearn.cross_decomposition.PLSRegression`

[14] Randall D. Tobias, *An Introduction to Partial Least Squares Regression*, SAS Institute Inc., Cary, 2011

[15] Hervé Abdi, *Partial Least Squares (PLS) Regression*, University of Texas, Dallas, 2003

[16] *Ammonium Sulfate*, National Institute of Standards and Technology (NIST), U.S. Secretary of Commerce, `https://webbook.nist.gov/cgi/cbook.cgi?ID=C7783202&Type=IR-SPEC&Index=1`, 2018