Saturated signals in spectroscopic imaging: why and how should we deal with this regularly observed phenomenon?

Alessandro Nardecchia, Vincent Motto-Ros, Ludovic Duponchel

Saturated signals in spectroscopic imaging: why and how should we deal with this regularly observed phenomenon?

Alessandro Nardecchia, [†] Vincent Motto-Ros,[‡] Ludovic Duponchel[†,*]

[†] Univ. Lille, CNRS, UMR 8516 – LASIRE – LAboratoire de Spectroscopie pour Les Interactions, La Réactivité et L'Environnement, Lille, F-59000.

[‡] Institut Lumière Matière, UMR 5306, Université Lyon 1 - CNRS, Université de Lyon, Villeurbanne, 69622, France.

ABSTRACT: We have all been confronted one day by saturated signals observed on acquired spectra, whatever the technique considered. A saturation, also known as clipping in signal processing, is a form of distortion that limits a signal once it exceeds a threshold. As a consequence, clipped or saturated bands with their characteristic plateau present numerical values that do not correspond to the analytical reality of the analyzed sample. Of course, analysts know that they cannot consider these erroneous values and therefore reconsider either sample preparation or instrument settings. Unfortunately, there are many experiments today (and this is the case in spectroscopic imaging) for which we will not be able to fight against the saturation effect that will undeniably be observed on the acquired spectra. The aim of this article is first to show why it is important to correct these saturation effects at the risk of having a biased view of the sample and more specifically in the context of multivariate data analysis. In a second step, we will look at strategies for managing saturated bands. An original concept will then be presented by considering saturated values as missing ones. A statistical imputation strategy will then be implemented in order to recover the information lost during the measurement.

Keywords: saturated signal, imaging spectroscopy, statistical imputation

INTRODUCTION

Saturation is a phenomenon regularly observed in spectroscopy. Its presence can be linked to various factors such as sample preparation, specific photon-matter interactions or even instrumental limitations in the detection chain. For example, too high extinction coefficients and/or too high pathlengths in the mid-infrared spectral range reduce so much the number of non-absorbed photons arriving at the detector that the absorbance levels are infinitely high, values that cannot of course be transcribed in a spectrum,

or by default in the form of a plateau. On the other hand, in the case of scattering or emission measurements as in Raman, fluorescence or LIBS (laser-induced breakdown spectroscopy), the number of photons collected by the measurement chain can be sometimes so important that it cannot be transcribed into the spectrum. Again in this case, a clipping effect is observed which does not allow to observe the real values to be measured. Generally speaking, we can say that a saturation may occur when a signal is recorded by a detector that has constraints on the range of data it can measure. This can therefore be the case when a signal is digitized using an analog-to-digital converter, or any other time an analog or digital signal is transformed, particularly in the presence of gain.

When a saturation phenomenon is observed on a spectrum acquired for bulk analysis of a single sample, the analysts know that they have has to reconsider the preparation of their samples or the acquisition parameters depending of course on the constraints related to the experiment under consideration. The newly acquired spectrum then has every chance this time to present values that are representative of the analytical reality of the sample. The situation is quite different when we have to do bulk analyses on a set of samples with the final objective of comparing their spectra. For this specific purpose, we must then set unique experimental conditions that will be applied to all samples. We could then easily observe perfectly exploitable spectra next to others that are potentially saturated. This is a situation that often occurs in spectroscopic imaging when exploring a single and complex heterogeneous sample. Indeed, for given acquisition conditions, hundreds, thousands or even hundreds of thousands of spectra are acquired in a region of interest of the sample. Since each spectrum corresponds to a specific micro-surface of the sample with a potentially different molecular distribution, it is quite likely that some of them are saturated. If we are lucky we might be able to find experimental conditions that remove these saturations. Nevertheless, we must not lose sight of the fact that it is not always possible to reproduce the experiment a second time, for example when the technique is destructive as in LIBS.

In general, we can say that we always try to avoid the saturation phenomenon as much as possible. Unfortunately, it is observed in many cases and it is necessary to deal with these data as they are. The question that then arises is the following: what should we do with saturated values that we know to be systematically erroneous? Figure 1a gives a schematic representation of a dataset with six spectra of which three contain saturations highlighted in red. We can notice first that it is not always the same bands that are saturated in this dataset used as a toy example. Second, the number of saturations in a given spectrum is quite variable. Figure 1b presents the two strategies typically used to manage potential spectral saturation in a dataset. The idea is finally very simple since knowing that saturated values do not represent the true values, it seems logical to remove them from the acquired dataset. We then have a first possibility which is to remove all the spectra as soon as they contain at least one saturated spectral variable also known as row-wise deletion. This strategy might seem satisfactory because it is simple to

implement, but it is not flawless. Indeed, we could then remove a spectrum made up of several hundreds of spectral variables and thus potentially a very large amount of molecular/atomic information just because a single variable would be saturated, for example. We would then have a significant loss of chemical information as in the present case study where only 50% of the spectra would be kept for multivariate analysis. In the specific case of spectroscopic imaging, we would then end up with areas of the image without defined chemical information. From a more statistical point of view, we would also have a biased analysis since we would no longer have the initial population of acquired spectra. In a second strategy known as column-wise deletion, we could suppress a spectral variable in the data set as soon as at least one of the spectra of the dataset presents a saturation on this same variable. This strategy is no more satisfactory because a significant loss of information would still be observed. In the case of the presented example, we notice that such a strategy would remove almost all the spectral information from the dataset. Thus even if these two strategies are regularly exploited in spectroscopy, we see that they are unsatisfactory on different aspects.

Starting from the observation that a saturated value in a spectrum is an erroneous one, we propose in this work to consider it as a missing value. It is indeed more relevant to say that a value could not be measured than to exploit a value that finally does not represent a reality. Thus in the matrix representation of the data set in Figure 1b, red boxes that were initially saturated values will become missing ones. In statistics, the art of dealing with missing values in a matrix is called imputation [1]. It is in fact the process of replacing missing data with substituted values. By approaching the problem of saturation in this way, we see that we can then work on a data set while keeping its initial dimensions, i.e. with the initial number of spectra and spectral variables resulting from the acquisition. Thus in this work, three different spectroscopic imaging datasets will first be used to show the need to manage saturations present in the spectra at the risk of seeing many artifacts during multivariate analyses generating biased chemical images and extracted spectral profiles. The principle of imputation will of course be explained and the analysis of the corrected datasets will allow us to demonstrate the benefits of this concept to find chemical images and corresponding spectroscopic information representative of the analytical reality of complex samples.

MATERIAL AND METHODS

Imputation

Imputation is a field of statistics. The great idea in imputation is to fill gaps in the data with plausible values, the uncertainty of which is coded in the data itself. There are many ways of doing data imputation today [1]. However, we will use in this work the so-called ‘multiple imputation’ now considered as the best general method to deal with incomplete data (i.e. containing missing values) in many scientific

domains [2–4]. Our goal here is of course not to redo a whole development of the theory of imputation but to explain some general principles in order to understand the results presented in this work. Readers who would nevertheless like to have all the details on this topic are invited to read other works specifically dedicated to statistic [1,3]. The two main approaches for imputing multivariate data are called joint modeling [5] (JM) and fully conditional specification (FCS), also known as multivariate imputation by chained equations (MICE) [6]. As the JM approach is often more constraining from a statistical point of view to be applied, the MICE method has been considered in this work. MICE specifies the multivariate imputation model on a variable-by-variable basis by a set of conditional densities, one for each incomplete variable (i.e. containing missing values). In this work, a regression model is developed using the complete variables of the matrix as input and a given incomplete variable as output. Once this model is established, we can then use it to predict missing values of spectra at this specific spectral variable based on known values in the matrix. We will thus have as many regression models developed as spectral variables containing missing values in the considered dataset. At the end of this imputation procedure, we then find a full matrix free of missing values that we can then explore with usual multivariate methods. In this work, all imputation calculations have been done under the R environment using MICE, an open source R package. Source code and documentation can be found at https://github.com/amices/mice.

Multivariate data analysis

Principal component analysis (PCA) is one of the most flexible and effective chemometric method for exploratory data analysis applied to hyperspectral imaging [7]. It is indeed very sensitive and thus allows the detection of very low variance levels. Its use will first of all allow to see artifacts generated during a direct use of saturated spectral datasets but also to estimate in a second time the efficiency of the corrections brought by our imputation strategy. All PCA calculations were performed under the Matlab 2016b environment.

Dataset #1: a simulated sample

The first hyperspectral imaging dataset consists of synthetic spectra that could have been acquired using LIBS. The advantage of using such simulations lies in the fact that all the parameters potentially influencing a given problem are under controlled. In this way, we often have a less biased view of the phenomena and a real generalization is possible. As we will see further on, it will also be a way to vary the importance of the saturation phenomenon. On the basis of the spectroscopic information given by the Kurucz database [8], we first simulated the emission spectra of silver, aluminum and arsenic by considering a typical plasma temperature and electron density (9000 K and 5.1016 cm^{-3} respectively). In order to be the most faithful with the spectral reality, we then applied a Lorentzian profile with a linewidth of 0.15 nm to each emission line, corresponding to the resolution of classical spectrometer used

in LIBS [9]. In the considered spectral range (250-350 nm), several emission lines of Ag, Al, and As were observed with various intensity ranges. On the basis of these three pure spectra, it was then possible to generate by linear combination 62880 spectra of mixtures in percentages ranging from 100 to 0 for each of them. A white noise of 5 % has also been added to each spectrum. In this way, we obtained a hyperspectral data cube defined by 131 pixels x 480 pixels x 2018 wavelengths. Figure 1S in supplementary material presents the three element spectra, all the generated spectra of mixtures in overlay mode as well as the spatial distribution of the different elements in this synthetic sample.

Dataset #2: a lung biopsy

The second dataset used in this work corresponds to a LIBS imaging experiment conducted on a lung biopsy of a patient with severe emphysema [10]. Note that the patient signed informed consent, and the clinical procedure was approved by the local ethics committee. The LIBS imaging has been conducted with a protocol dedicated to paraffin-embedded tissues, as described in a previous work [11]. The aim of such application was to characterize the distribution of metallic particles (from nanometric to micrometric size) in tissue biopsies, which represent a precious help for clinicians to diagnose the cause of the exposition (i.e. environmental and/or occupational). This spectroscopic experiment is a good example of a case where saturated spectra cannot be avoided. Since the concentration, composition, location and size of the particles are not known prior the experiment, the measurement system requires an extremely large dynamic in term of detection, typically from few ppm to a few percent in mass. Despite our efforts to set up the experimental parameters as optimized as possible, it is not uncommon to have a significant number of spectra showing saturations on a LIBS image as in this case. The size of the analyzed area of the biopsy was 5.42 mm long by 3.18 mm wide with a spatial resolution of 20 µm. We have thus acquired 43089 spectra over a spectral range from 282.01 to 310.03 nm and an approximate spectral resolution of 0.04 nm. The hyperspectral data cube was therefore defined by 271 pixels x 159 pixels x 644 wavelengths.

Dataset #3: a rock section

The third dataset was also acquired using a LIBS imaging instrument on a banded iron formation rock consisting of alternating layers of iron oxides and silicates. We selected this sample because it somehow allowed us to find approximately the same chemical distributions for two contiguous zones of the sample on which we could set different acquisition parameters. This procedure of selection of analysis area was necessary because we must not forget that LIBS is a destructive technique. Therefore, we could not analyze the same area several times. As a consequence, two successive zones each having a size of 20 mm long and 2 mm wide were analyzed following the protocol already used for other work [12]. A spatial resolution of 20 µm and a spectral range from 245.85 to 334.03 nm were considered for

these acquisitions generating two hyperspectral datasets defined each by 1000 pixels x 100 pixels x 2048 wavelengths. The laser pulse energy and the detection gate were adjusted for these two sub-zones in order to control the saturation level. Indeed, our aim was to obtain no saturated lines on the first zone of the sample (gate: 1 µs; energy: 1.2 mJ) and saturation of Si emission lines on the second zone (gate: 5 µs; energy: 1.2 mJ). All the other acquisition parameters such as the delay and detector gain was kept constant.

RESULTS AND DISCUSSION

Before tackling the problem of correcting saturation in the spectra, it is important to understand how this is necessary to manage it, at the risk of giving a completely biased vision of the analyzed sample. For this purpose, we will use the first dataset of simulated spectra. Figure 2a thus presents the results of a first principal component analysis applied to the raw data (i.e. without saturated signals). Unsurprisingly, we note first of all that there are three significant eigenvalues in the scree plot that correspond to three spectral contributions. More specifically, the first three principal components respectively extract the spectra of the three pure elements Ag, Al and As. This is quite logical since there is no correlation between these elements in the considered dataset. The scores images then perfectly reproduce the distributions of the three elements given in SI Figure 1S. In a natural way, the fourth principal component extracts the noise variance. From the raw data, we then simulate a first level of saturation by clipping all emissions above 30, knowing that the maximum emission observed on the initial spectra is around 43. SI Figure 2S gives the location of the saturated signals at the spatial and spectral levels. 5047 spectra thus present saturations, i.e. 8% of all the spectra. We notice that saturations are present in areas where silver is the most concentrated with a percentage higher than 80%. From a spectral point of view, it is the most intense line of silver which is naturally saturated. Figure 2b shows again the PCA results on these new saturated data. The consequences are not long in coming since a fourth significant component is already detected in the eigenvalues scree plot. Of course, this is not normal because we know that only three elements are present. Compared to the initial results (in Figure 2a), both the first principal component and the first scores map are no different. Nevertheless, there is a small decrease in the expressed variance from 21.42 to 20.44%. As far as the second and third scores maps are concerned, they are quite comparable to those observed from the non-saturated dataset. On the other hand, the corresponding principal components show small artifacts in the spectral region of the saturated Ag band (highlighted by red boxes in the corresponding figure). Finally, the fourth principal component specifically reflects the saturation phenomenon observed on silver for an expressed variance of 0.08%. We see many structures in the corresponding scores map but we know that they do not reflect any analytical reality. We observe even more the typical W-shaped artifact in this principal component. This shape can

be explained quite well. Indeed, when saturations are potentially present at a given wavelength, principal components have to express the variance precisely at this wavelength for the unsaturated spectra in the dataset but other ones also have to express variances specifically localized on the feet of this same peak. In other words, a clipped band at a certain wavelength of a given spectrum is no longer homothetic to an unsaturated band of another spectrum. So if we do not have any prior information about this dataset, we see that even a limited saturation level can induce the extraction of erroneous information at the spatial and spectral levels about the sample being explored. It is then interesting to amplify the saturation effect by considering this time a saturation level equal to 20. Under these new conditions, 10119 spectra are saturated, i.e. 16% of all the spectra. Once again, saturations are present in areas where Ag is the most concentrated, but this time for values higher than 50% (SI Figure 3S). We are also starting to see saturated spectra for pure Al pixels. Figure 2c show PCA results of this new dataset. Four significant contributions are still observed, but with an even greater influence of artifacts on all components. We notice thus on the first principal component which should be specific to Ag that contributions from Al and As are now easily observed. We note here that the presence of saturations can also create spurious correlations. The second and third components have even greater expressed variances and ever more pronounced 'W-shapes'. While the scores maps are relatively little changed under these new conditions for the first three principal components, this is not at all the case for the fourth one. This high-contrast, low-noise scores image could indeed lead us to believe that real chemical compounds are present, which is of course not the case. In a final step, the saturation is further increased by considering this time a level equal to 10. This situation is extreme since 27870 spectra are now saturated, i.e. 44% of the dataset. What is more, the saturated pixel location map shows that this percentage is even underestimated since almost all of the areas that should contain the three elements are almost all saturated, the unsaturated areas being mainly the background (SI Figure 4S). At the same time, we observe that almost all emission lines show saturation over the entire spectral range. PCA results of this new dataset is given in Figure 2d. Under these conditions where saturation is omnipresent, six significant contributions are now detected. The first principal components are more and more perturbed. They are now undeniably different from pure spectra extracted on unsaturated spectra. As examples, the first principal component contains distinct contributions from all three elements and the following ones, which contain more and more artifacts, have equally increasing explained variances. Two new principal components 5 and 6 are also extracted in these conditions with quite singular scores maps. The presence of these additional principal components is explained by the fact that the variance of all the saturated peaks must of course be explained, but also that they are not necessarily saturated at the same time in all the spectra of the dataset. Generally speaking, we can say that the more spectral variables containing saturations, the more parasitic principal components and biased scores maps are extracted. From this first experiment, it is

obvious that we cannot directly process saturated spectra with multivariate tools at the risk of making very hazardous exploration of unknown and complex samples for which we have no a priori. Based on this observation, we know that we must now absolutely manage these saturations. Thus, if we wanted to implement a row- or a column-wise strategy that is simple to set up in order to eliminate these saturations, we would quickly observe too many deleted pixels or a particularly small explored spectral domain. It is in this sense that the proposed imputation strategy makes sense by first considering saturated signals as missing values and then applying the MICE approach to make statistical estimates of the latter, i.e. retrieve lost spectral information and consequently a full data matrix. Imputation was therefore applied to the previous datasets by considering the three levels of saturation. In order to appreciate the quality of the data reconstruction, we simply reapplied PCA on these three new datasets. Figure 3 shows the results concerning the intermediate saturation level equal to 20. The results for the other levels are presented in the supplementary material (SI Figure 5S). By comparing these new extractions with those obtained on unsaturated data, we observe rather spectacular results. First of all, we recover the three significant components on the eigenvalues scree plot, which is consistent with the initial results. Moreover, principal components and corresponding scores maps are also very comparable to the initial ones. These good results can be explained by the fact that the multivariate regressions used in the MICE approach predict missing values rather well. By way of illustration, Figure 4 shows the emission predicted by the imputation model as a function of known values at the wavelength 323.96 nm from non-saturated data, this silver emission line being the most often saturated for a saturation level equal to 20. Looking specifically at the results concerning the most saturated dataset (SI Figure S5, saturation level equal 10), some readers might say that despite the three significant contributions detected on the scree plot, it is possible to observe information related to a fourth component at both spectral and spatial levels. This would be quite commendable but we must not lose sight of the fact that these contributions are very close to the noise level. Moreover, these results were obtained from a dataset for which almost all the spectra were saturated, which could not be more challenging.

In this second part, we propose to explore a lung biopsy sample. This sample is particularly interesting because the analyzed area of lung presents a certain diversity of materials since we have naturally biological tissues but also mineral phases and metal particles localized in specific sub-areas. In these conditions, we quickly understand that it is almost impossible to find acquisition conditions allowing us to avoid saturation over the entire surface analyzed. In a way, one always make a bet before such an analysis because the considered spectroscopy is destructive and it is not possible to return to this sample area with new acquisition parameters. Location of the spectra containing saturations on the surface of this sample is shown in SI Figure 6S. Although only 1014 spectra out of the 43089 in total are saturated (i.e. 2.35%), this phenomenon is finally observable almost everywhere, mainly on two well-localized

areas (denoted A and B in this figure) but also in the form of single pixels scattered over almost the entire surface of the sample. Additionally, Figure 6S shows that saturations are observed for almost all emission bands in the considered spectral range. Figure 5a and b shows PCA results on raw spectral data and spectra corrected with imputation respectively. Differences are noticed very quickly if we look at the contributions of each principal component in these two conditions two by two. So even though the first principal component is quite comparable in both cases with the main spectral contributions observed for Mg and Si but also smaller ones for Al and Fe, the associated scores maps are very different. Indeed, there is an overestimation of this first contribution for raw data on zones A and B of the sample but also widely around zone B. For its part, the first scores map associated with the corrected data mainly locates this contribution on the periphery of zones A and B or on specific pixels scattered outside these zones. Another way to observe these differences is to compare the histograms of positive scores for this first component in the two conditions (SI Figure 7S). The saturation effect thus limits the range of scores values that should be observed and profoundly changes the structure of the distribution and therefore the visual perception one might have of it. For the second component, we are in much the same situation as before. We therefore have very comparable second principal components on the raw and imputed data. The Mg contribution is now anticorrelated to the Si, Fe and Al ones. On the other hand, once again there are differences on the scores maps for this component in the two conditions. Negative scores (blue color scale) are thus distributed more homogeneously in areas A and B when the spectral data are imputed. It is from the third principal components that we observe the largest spectral differences between the two conditions. Thus for raw data, typical W-shaped artifacts are observed (in red in Figure 5) around the Mg contribution with correlations or anticorrelations with other elements. We observe on this occasion that the third and fourth principal components are extracted from raw data to express the saturation of pixels mainly located in the B zone of the sample. Even more specifically, we can see on the third principal component that the W-shaped artifact on Mg is positively correlated with another Ca contribution around 300 nm. This component thus just testifies to the simultaneous saturation of emission bands associated with the Mg and Ca elements on specific pixels according to the information given in SI Figure 6S. This example shows a very good example of spurious correlation created by the saturation phenomenon, which no longer exists once the data are corrected by imputation. Finally, the imputation strategy allows the appearance of dispersed particles opposing the Si and Al elements for the third principal component and the Ti et Al elements for the fourth one. It is obvious that such potentially less biased observations of particles represent a precious help for clinicians to diagnose the causes of the patient's exposure. From a general point of view, it is very interesting to see how a small percentage of saturated spectra can have an influence on a multivariate exploration method as sensitive as principal component analysis. This experience shows again here the necessity not to neglect

the saturation phenomenon by setting up an adapted correction method such as imputation prior any chemometric analysis.

This last part of this work is dedicated to the analysis of the rock sample. As a reminder, two contiguous regions of the sample were analyzed considering two acquisition settings. In this way, we analyzed the sample by ensuring the absence of saturation for a first area but also its presence in the second one. Figure 8S shows the location of pixels containing saturations on the surface of the second sample area. It can be said that in this case saturation is omnipresent since it is observed on 32.5 percent of the analyzed surface. On the other hand, the same figure shows that this time these saturations are only found on the specific contribution of an element, namely silicon around 288 nm. Figure 6 presents the three principal component analysis calculated on the first sample area (i.e. with no saturation), on the raw data of the second sample area (i.e. with saturations) and on the imputed data (i.e. corrected ones) of the same area. By comparing the principal components two by two in Figures 6a and 6b, we observe very quickly the impact of saturation since we find the typical W-shaped artifact around 288 nm for components 3 and 4. The situation is even more critical for the fifth principal component with completely different profiles between Figures 6a and 6b. In fact, it is above all the saturation effect that is expressed here for the second area of the sample. Finally, by comparing the results on the imputed data in Figure 6c and the unsaturated data of the first sample area, we observe a perfect agreement between extracted profiles demonstrating the capacity of our approach to correct the saturated spectral data.

CONCLUSION

As we have seen in the work, it is crucial to consider the phenomenon of saturation present in the spectra. Through different datasets we have indeed shown that its presence quickly induces artifacts on spectral profiles but also on generated images when multivariate tools are used for their exploration. Make no mistake, even the presence of a limited percentage of saturated spectra in a given dataset can have an impact on the veracity of the chemometric results. It is obvious that it is absolutely necessary to avoid the presence of saturation in the acquired spectroscopic data whenever possible by modifying, for example, the sample preparation or the acquisition parameters. Unfortunately, there are many situations where this phenomenon is observed as in LIBS imaging and we have to find solutions to exploit these acquired data anyway. The usual column- or row-wise deletion is not a satisfactory solution because it can be accompanied by a large loss of spectral information in the dataset. As a consequence, we would have a partial or even biased view both at the spectral and spatial level of the sample. All the originality of our work was to consider the saturated signals as values that had not really been measured and by extension as missing values. The goal being to preserve all the spectral and spatial dimensions of the dataset, statistical imputation allowed us to retrieve complete data cubes consistent with the analytical

reality of the samples considered as shown in the results. With this new approach, we will potentially have a chance to explore all those datasets that we think are being lost due to saturated signals.

AUTHOR INFORMATION

Corresponding Author

(*) Corresponding author: ludovic.duponchel@univ-lille.fr

ORCID

Alessandro Nardecchia: 0000-0003-3131-8436

Vincent Motto-Ros: 0000-0001-6063-5532

Ludovic Duponchel: 0000-0002-7206-4498

Author Contributions

The manuscript was written through contributions of all authors.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENT

The authors would like to acknowledge the French ANR research program for their financial support. In addition, the authors are grateful to B. Busser, V. Bonneterre and G. Panczer for helpful discussions.

REFERENCES

[1] S. van Buuren, Flexible Imputation of Missing Data, Chapman and Hall/CRC, 2012. https://doi.org/10.1201/b11826.

[2] F. Scheuren, Multiple Imputation: How It Began and Continues, The American Statistician. 59 (2005) 315–319. https://doi.org/10.1198/000313005X74016.

[3] D.B. Rubin, ed., Multiple Imputation for Nonresponse in Surveys, in: Wiley Series in Probability and Statistics, John Wiley & Sons, Inc., Hoboken, NJ, USA, 1987: pp. i–xxix. https://doi.org/10.1002/9780470316696.

[4] D.B. Rubin, Multiple Imputation After 18+ Years, Journal of the American Statistical Association. 91 (1996) 473–489. https://doi.org/10.2307/2291635.

[5] J.L. Schafer, Analysis of Incomplete Multivariate Data, Chapman and Hall/CRC, 1997. https://doi.org/10.1201/9780367803025.

[6] S. van Buuren, K. Groothuis-Oudshoorn, mice: Multivariate Imputation by Chained Equations in R, Journal of Statistical Software. 45 (2011). https://doi.org/10.18637/jss.v045.i03.

[7] S. Moncayo, L. Duponchel, N. Mousavipak, G. Panczer, F. Trichard, B. Bousquet, F. Pelascini, V. Motto-Ros, Exploration of megapixel hyperspectral LIBS images using principal component analysis, J. Anal. At. Spectrom. 33 (2018) 210–220. https://doi.org/10.1039/C7JA00398F.

[8] R. Kurucz, B. Bell, Atomic Line Data, Atomic Line Data (R.L. Kurucz and B. Bell) Kurucz CD-ROM No. 23. Cambridge, Mass.: Smithsonian Astrophysical Observatory, 1995. 23 (1995). http://adsabs.harvard.edu/abs/1995KurCD..23.....K (accessed October 7, 2020).

[9] V. Motto-Ros, S. Moncayo, F. Trichard, F. Pelascini, Investigation of signal extraction in the frame of laser induced breakdown spectroscopy imaging, Spectrochimica Acta Part B: Atomic Spectroscopy. 155 (2019) 127–133. https://doi.org/10.1016/j.sab.2019.04.004.

[10] B. Busser, V. Bonneterre, L. Sancey, V. Motto-Ros, LIBS Imaging Is Entering the Clinic as a New Diagnostic Tool, Spectroscopy. 35 (2020) 29–31.

[11] S. Moncayo, F. Trichard, B. Busser, M. Sabatier-Vincent, F. Pelascini, N. Pinel, I. Templier, J. Charles, L. Sancey, V. Motto-Ros, Multi-elemental imaging of paraffin-embedded human samples by laser-induced breakdown spectroscopy, Spectrochimica Acta Part B: Atomic Spectroscopy. 133 (2017) 40–44. https://doi.org/10.1016/j.sab.2017.04.013.

[12] J.O. Cáceres, F. Pelascini, V. Motto-Ros, S. Moncayo, F. Trichard, G. Panczer, A. Marín-Roldán, J.A. Cruz, I. Coronado, J. Martín-Chivelet, Megapixel multi-elemental imaging by Laser-Induced Breakdown Spectroscopy, a technology with considerable potential for paleoclimate studies, Sci Rep. 7 (2017) 1–11. https://doi.org/10.1038/s41598-017-05437-3.

FIGURES

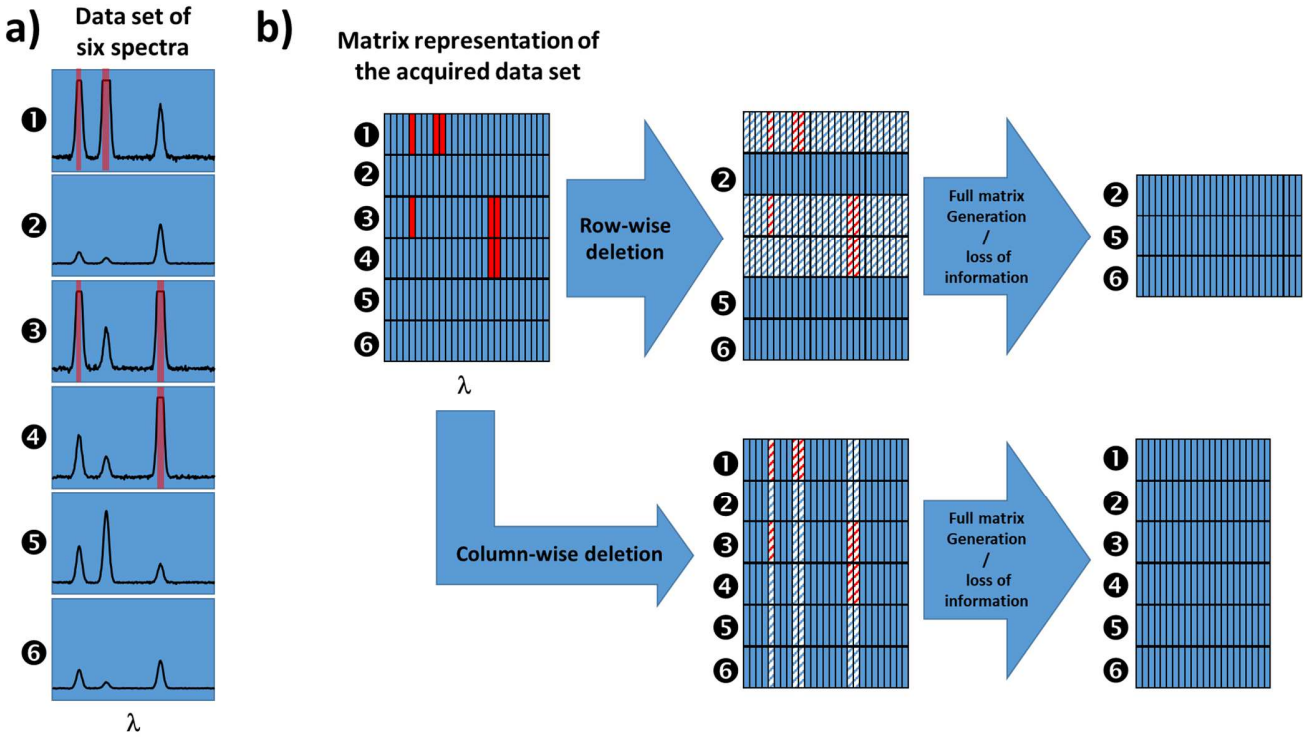


Figure 1. a) A schematic representation of a toy example with six spectra containing saturations highlighted in red. b) The two conventional strategies to manage saturated signals in a dataset.

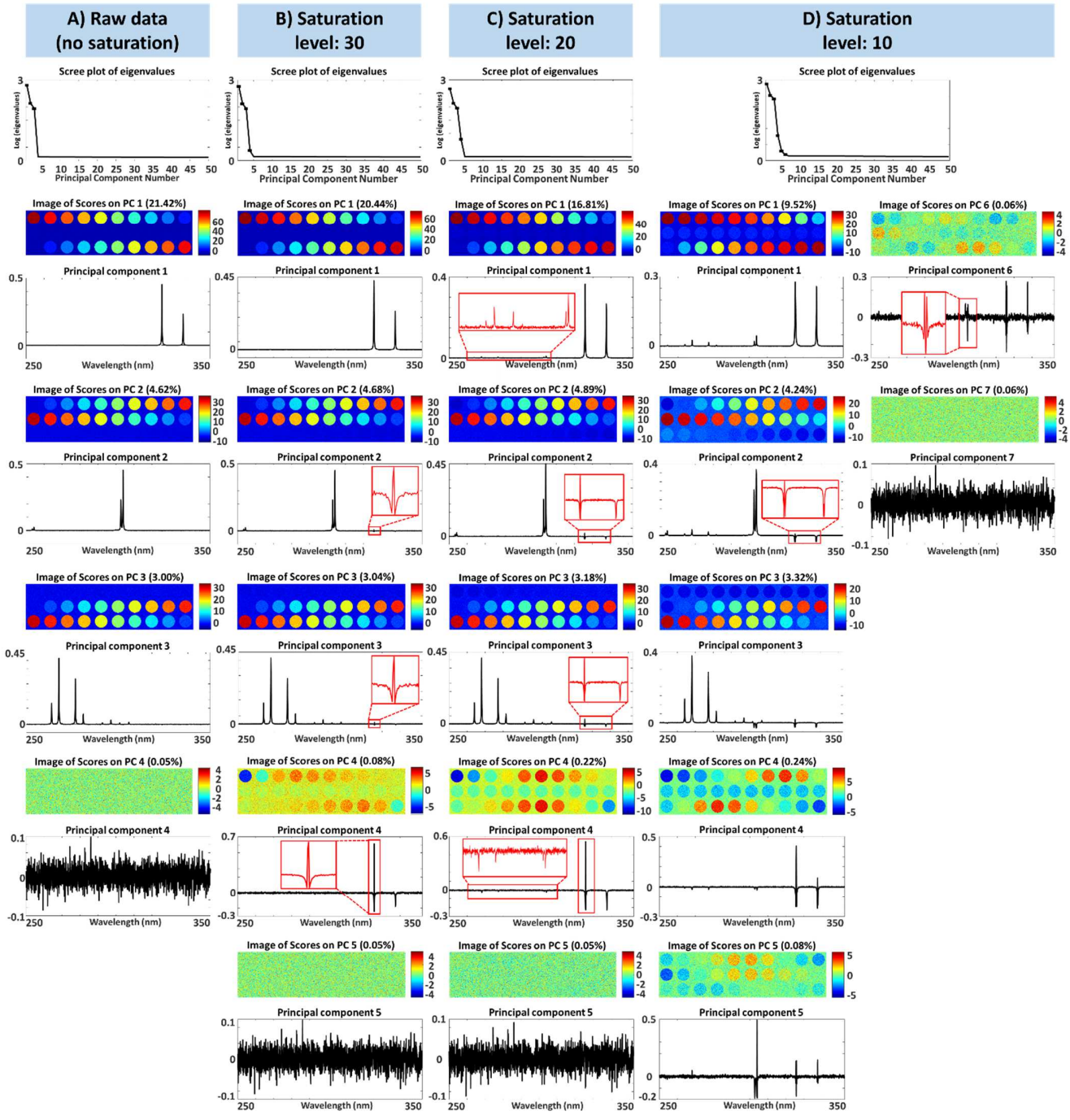


Figure 2. Principal component analysis on a) raw data, b) on the saturated dataset with a 30 saturation level, c) with a 20 saturation level and, d) with a 10 saturation level.

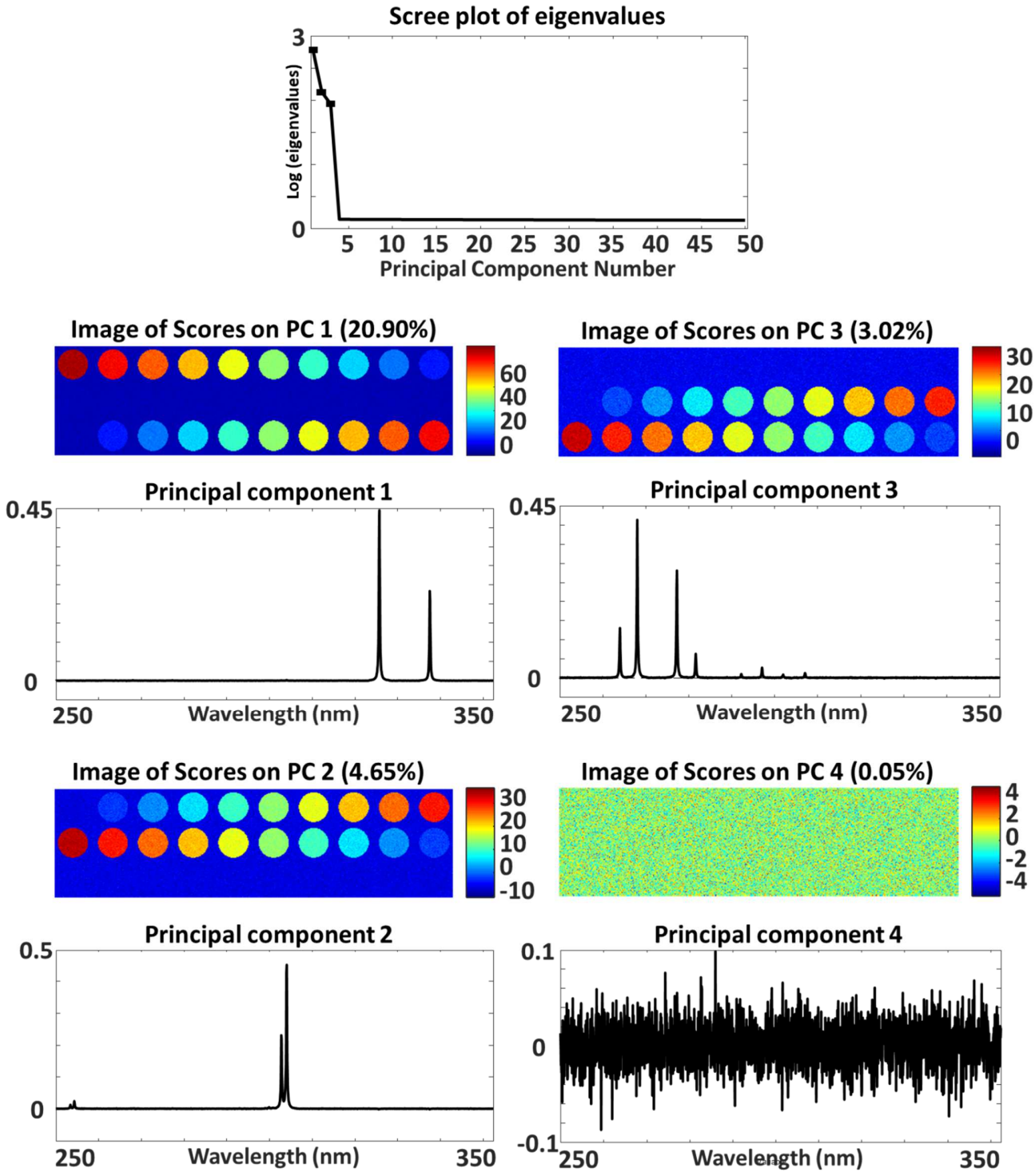


Figure 3. Principal component analysis on the imputed dataset with an initial saturation level equal to 20.

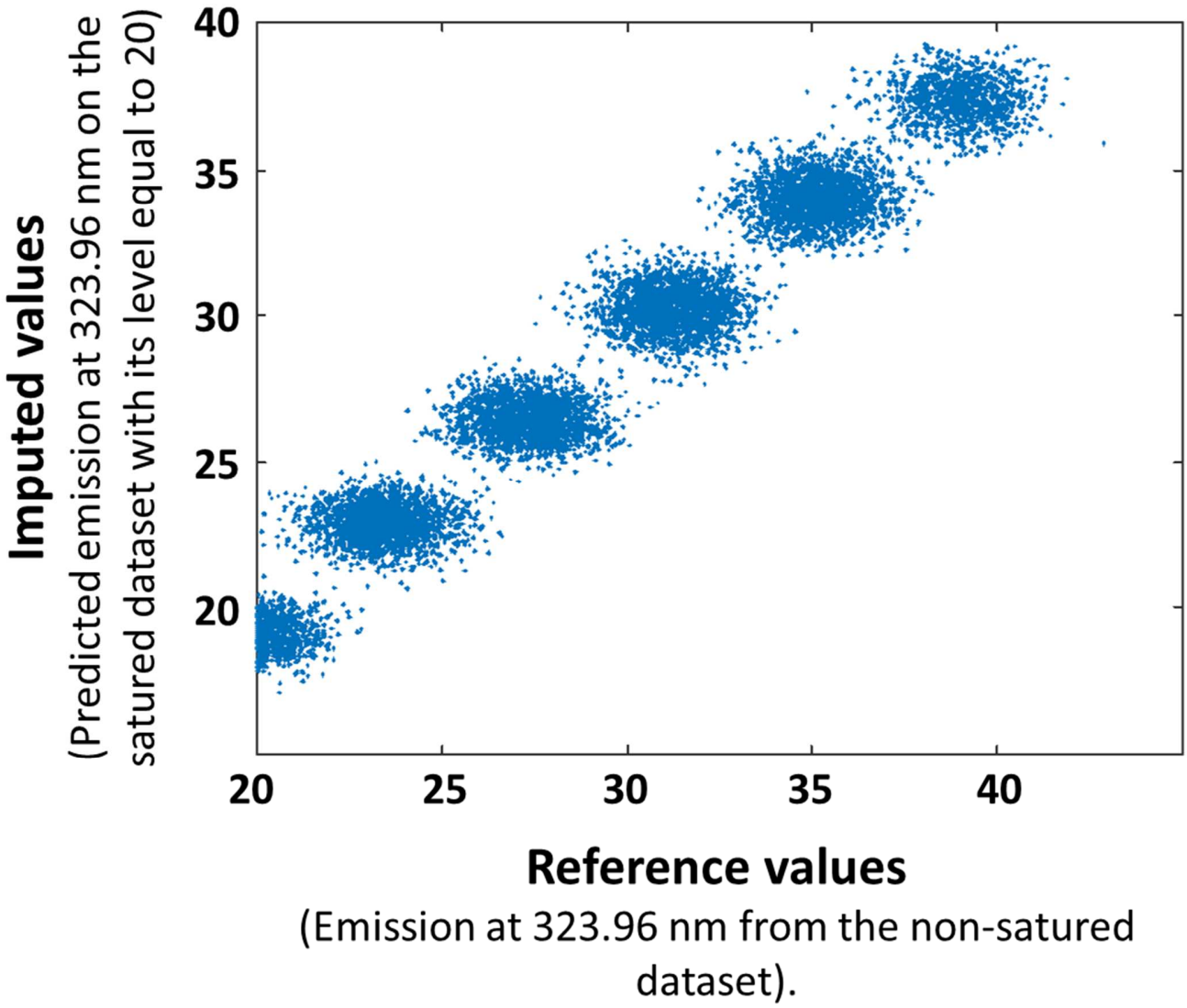


Figure 4. Emission of a silver line predicted by the imputation model as a function of known values at the wavelength 323.96 nm from non-saturated data.

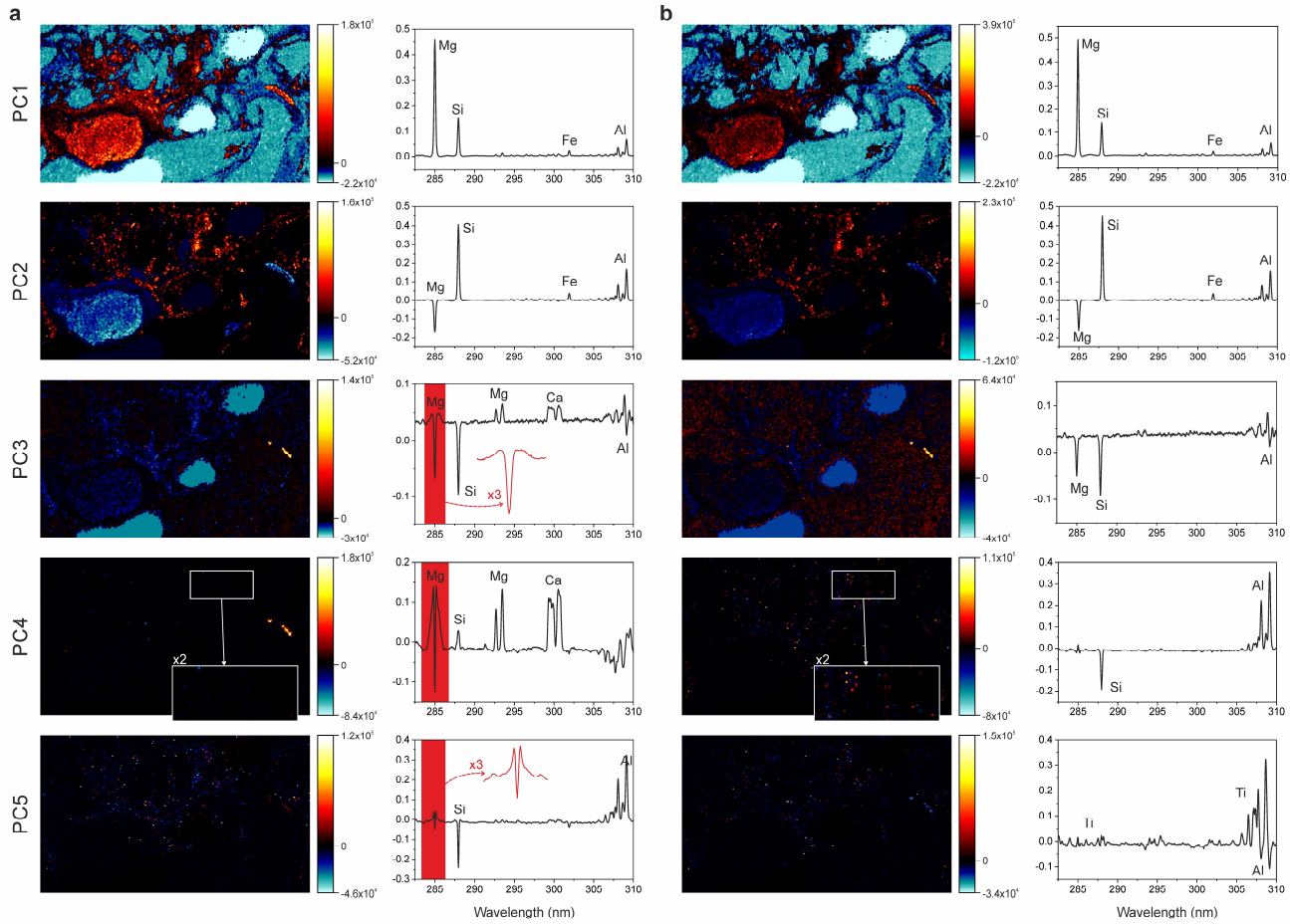


Figure 5. a) Principal component analysis on raw spectral data of the lung biopsy. (b) Same analysis on spectra corrected with imputation.

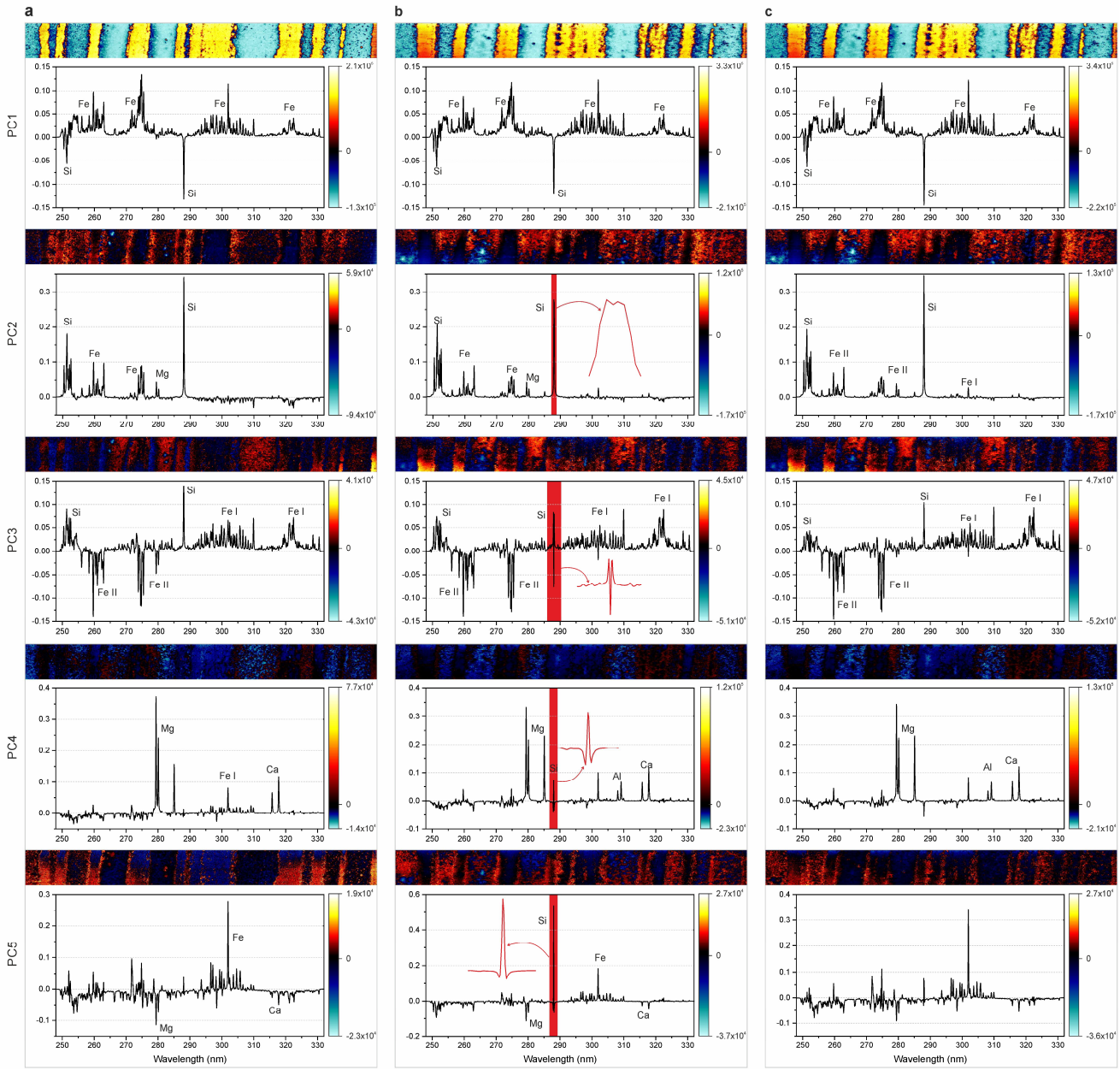
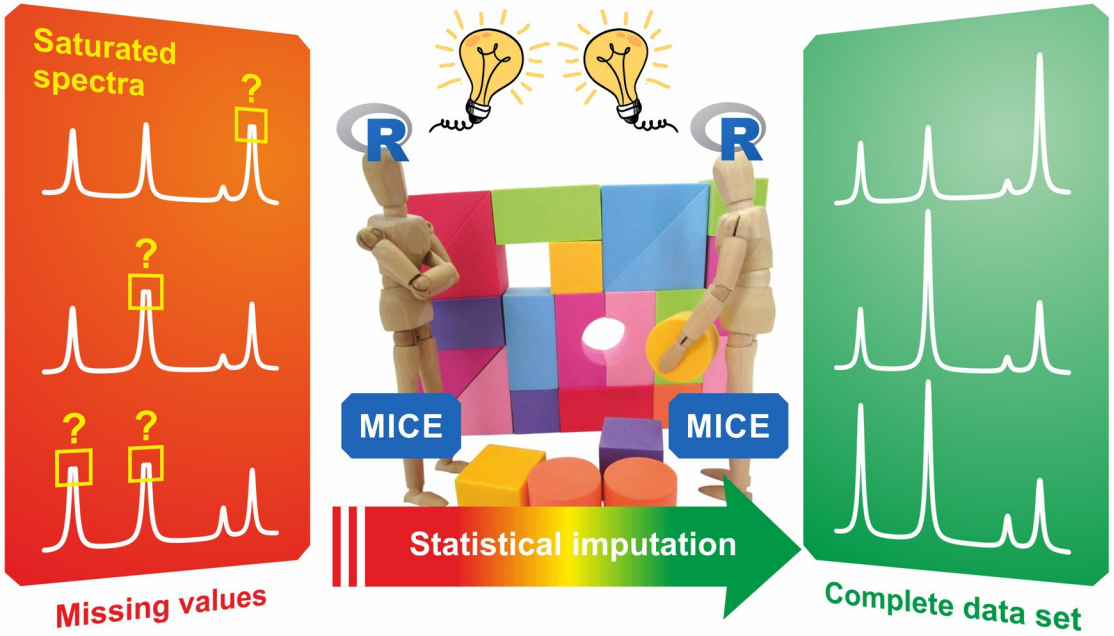


Figure 6. Three principal component analysis calculated on, a) the first sample area (i.e. with no saturation), b) the raw data of the second sample area (i.e. with saturations) and c) the imputed data (i.e. corrected ones) of the same area.


Saturated spectra
?
?
? ?
Missing values
R
R
MICE
MICE
Statistical imputation
Complete data set