Using Data Science to Improve the Identification of Plant Nutritional Status

David Condaminet
Normandie Univ, UNICAEN, ENSICAEN, CNRS
UMR GREYC - 14000 Caen, France
david.condaminet@unicaen.fr

Albrecht Zimmermann
Normandie Univ, UNICAEN, ENSICAEN, CNRS
UMR GREYC - 14000 Caen, France
albrecht.zimmermann@unicaen.fr

Bastien Billiot
Centre Mondial de l'Innovation Roullier
Laboratoire de Nutrition Végétale
35400 Saint-Malo, France
bastien.billiot@roullier.com

Bruno Crémilleux
Normandie Univ
UNICAEN, ENSICAEN, CNRS
UMR GREYC - 14000 Caen, France
bruno.cremilleux@unicaen.fr

Sylvain Pluchon
Centre Mondial de l'Innovation Roullier
Laboratoire de Nutrition Végétale
35400 Saint-Malo, France
sylvain.pluchon@roullier.com

***Abstract*—Developing products for improving plant nutritional status, e.g. fertilizers or plant growth regulators, is an important topic to move towards sustainability in agriculture and to ensure to feed the world population. A key challenge is to identify *when*, *what*, and *how much* nutrients to add to plants' growth environment. In this paper, we study a use case on how to characterize rapeseed plant nutrient deficiencies during their growth. A promising approach consists of deriving data from spectroscopy of leaves, and using this representation to predict what kind of deficiency (if any) plants are undergoing. We are considering three research questions: 1) from which day after onset of a nutrient deficiency we can identify it, 2) whether leaves that have sprouted under nutrient-rich conditions can still help in identifying problems. Third, and most importantly, performing the spectroscopy on the full range of wavelengths is expensive, which under production conditions allows for only relatively few samples. We therefore explore how to perform dimensionality reduction and preprocessing to achieve good predictive accuracy. We show that 1) deficiencies can be identified early on, 2) leaf generations help to predict nutrient deficiencies, and 3) that preprocessing increases the accuracy and dimensionality reduction can be performed without loss of accuracy. Along the way, we find that our some of our industry partners' assumptions about the data do not seem to be borne out by our empirical results, and that the subset of data they initially selected turns out to be too easy to model. The full data leads to more informative insights.**

***Index Terms*—agronomy, plant nutrition, dimensionality reduction, subgroup discovery, classification**

I. Introduction

Agricultural chemistry is concerned with exploring the biochemical processes governing plant (or animal) growth, maturation, diseases etc. The end goal of such research is the development of substances that can support or control processes, e.g. fertilizer, herbicides and pesticides, or plant growth regulators. Even though the field has been an important contributor to the "green revolution" of the 20^{th} century, it has acquired renewed relevance recently since agricultural paradigms are shifting.

In 21^{st} century agriculture, and in particular in the context of plant nutrition, both how to affect improvements in soil quality, and how to ensure the sustainability of the intervention need to be considered. In addition, production yields should be enough to feed an expected 10 billion people by 2050. To this end, bringing out plant nutrients should follow the maxim: "as much as necessary, as little as possible". One step towards this goal consists of identifying precisely *when*, *what*, and *how much* of nutrients to add to plants' growth environment. Adding too little or too late will stunt growth (and yields), whereas adding too much or too soon puts pressure on ecosystems and ground water (and might also stunt growth and yield).

This paper presents joint work between the GREYC Lab (Caen, France) and the Centre Mondial de l'Innovation Roullier (Saint-Malo, France). The purpose of the Centre Mondial de l'Innovation Roullier is to accelerate the development of advanced technical products to respond to the needs for crops in an increasingly constrained economic, social and environmental context. Increasing their innovation capabilities necessarily involves developing their product screening capabilities, both in terms of frequency and characterization parameters. It is this context that the study presented in this paper fits into.

A widely used method (cf. Section II) of identifying stresses on plants consists of performing spectroscopic analysis of plant leaves. Our industry partners approached us with experimentally generated data and three research questions:

1) From which day after onset of a nutrient deficiency can it be reliably predicted?
2) Which leaf generations (a concept that we will explain later on) are useful for predicting such deficiencies?
3) How can one reduce the set of wavelengths to be measured?

The latter point requires some clarification: a lack of knowledge of the most informative wavelengths requires the use of

978-1-7281-8206-3/20/$31.00 ©2020 IEEE
DOI 10.1109/DSAA49011.2020.00064

wide-range spectrometers, which is a) costly, and b) degrades the measurement precision for each wavelength (wavelength for short). Related to this, spectroscopic measurements are only a means to an end, and feature reduction in the sense of removing uninformative or redundant features could aid eventual classification methods to better predict the stress a particular plant undergoes.

This is an *exploratory* paper in the sense that we first state the problem setting (Section III), characterize the data in Section IV, and perform an initial analysis of the spectroscopic measurements data to get a better understanding (Section V).

In Section VI-A, we discuss the results of data treatment methods proposed by our industry partners, informed by their own experience and the state of the literature (Section II), before analyzing the data in more detail, showing the impact of the leaf generations (Section VI-B) and the date of transplantation (Section VI-C) as well as different techniques to improve the data representation for characterizing deficiency conditions (Section VII). Regarding our research questions, we show that 1) deficiencies can be identified early on, 2) all leaf generations help to predict nutrient deficiencies (including $L1$ leaves which was unexpected for our industry partners), and 3) that preprocessing can increase the accuracy and the number of wavelengths can be divided by 3 without loss of accuracy. Another contribution validates our partners decision to invite our expertise into the process: we find that some assumptions of our industry partners about their data do not seem to be borne out by our empirical results and that working on the full dataset leads to more informative insights.

II. Related work

While machine learning and data mining *has* been used in connection with agrochemical products, it has been in the context of fertilizer recommendation, disease and weed detection, field management, or risk assessment. For a survey in agriculture see [1].

Dimitriadis *et al.* propose to use machine learning methods to make decisions regarding irrigation, fertilizer, and pesticide application [2]. Kuchenbuch *et al.* try to predict yield increases based on soil and plant characteristics [3]. Yu *et al.* use an ensemble of neural networks to address the same problem [4]. Recently, Shekoofa *et al.* evaluated techniques that not only predict maize yield but also aid in understanding it [5], settling on decision tree models. Yuan *et al.* used Gaussian processes to optimize fertilizer application [6].

A work that is very closely related to our proposal for identifying deficiency conditions used SVMs and ANNs to predict nitrogen and weed stress [7]. Their data set was much smaller than our, only 360 data points, and they report 81% classification accuracy for nitrogen stress and 89% for weed stress. That work was only interested in the feasibility of automatic classification, however, whereas we explore when and which data to acquire. The derived information should help calibrating data acquisition more carefully, and therefore aiding in the future application of data science methods.

Gholap *et al.* have proposed automatically predicting soil characteristics [8], a question that is obviously connected to how to improve its nutritional profile.

Jay *et al.* compare different hand-crafted *vegetation indices*, derived from spectral data, and parameters derived from inverting simulation models for identifying the nitrogen content of leaves [9]. While our samples were created under controlled conditions, their samples came from production plantings and they do not consider explicit stress conditions.

There is also work on selecting which bandwidths of spectral imaging to use for different tasks, often using partial least squares regression (PLSR) [10] or related methods, selecting wavelengths based on the strength of regression coefficients. Peng *et al.* performed prediction of salt concentrations in soil, using PLSR and training a support vector machine on the reduced representation. Li *et al.* and Guo *et al.* [11], [12] propose using a repeated subsampling (similar to bagging) to reject all wavelengths that are not consistently identified as informative, followed by a PCA-like procedure that selects features that allow for the largest projection into a plane orthogonal to the already selected features. They report that they can reduce the full descriptor space from 2151 wavelengths to 31, and further to 7 via multiple linear regression, and that this approach outperforms selection by PLSR. All such methods suppose that the entire spectrum is available to select from, however, and typically do not report specific wavelengths as informative. By contrast, the purpose of our work is to reduce the cost and increase the precision of the imaging system by selecting a subset of wavelengths beforehand to deploy detectors for.

Benoit *et al.*, finally, model photodetectors with Gaussian depending on the central wavelength and the bandwidth of the channel, select optimal parameters (for the discriminative task at hand) via Kullback-Leibler divergence, and report better results than PLSR or CDA. Their stated motivation is close to ours, limiting the amount of sensors that need to be deployed but their approach closer to the dimensionality reduction we describe in Section VII. The main difference is that the authors fix the number of "groupings" and search exhaustively for the best center and width.

III. The problem setting

The overarching goal is the automatic, precise, and early identification of nutrient deficiencies. The necessary information can be collected via spectroscopy but there are several questions that remained unclear to our industry partners who lack in-house data science expertise and therefore approached us.

First, it is not clear how early such deficiencies can actually be detected. Early detection is preferred since it requires less intervention to correct the problem.

Second, nutrient deficiencies will be expected to arise *during* the growth process, and while they should be easy to predict from strongly deficient leaves, we want to know whether leaves that experience nutrient stress *after* they have sprouted are also informative.

Finally, it is important to identify which wavelengths are informative, and using the insight to tailor the sensor setup. Wide-range spectrometers, like the one used to generate the data in this paper, used by laboratories specialized in this kind of analysis are expensive (∼45k €). This price point makes the acquisition of several of them prohibitive, and therefore slows data acquisition.

The goal of our analysis is therefore to reduce the range of wavelength needed to help with making data acquisition cheaper.

Several smaller setups of fewer sensors focused on particular wavelengths could be deployed at a lower price, generating more (and more precise) data. As we will see in our preliminary study (cf. Section V), even leaf samples from the same generation and experiencing the same conditions show differing wavelength reflectance values. This range could be tightened with smaller setups. While we discuss all three questions in our work, the biggest part is taken up with the third one, which requires more extensive experimentation.

IV. The data

Our industry partners have provided us with data of rapeseed plants that have been put into situations of nutrient stress in company greenhouses. These data (called FD for "Full Dataset") are generated in the following manner:

1) Plants are seeded into soil.
2) Plants are transplanted into a hydroponic solution containing all nutrients.
3) A week after the transplant, the hydroponic solution is replaced with a nutrient-deficient one.
4) Every three or four days, the hydroponic solution is replaced, and spectrographic analysis is performed (see below).

Plants belong to eight different classes and the same experimental conditions are applied during the experimental cycle except the nutrients:

- (1) FULL – no nutrient stress (control group)
- complete lack of (2) nitrogen (N0), (3) sulfur (S0), or (4) phosphorus (P0), respectively
- partial lack of (5) nitrogen (NINT), (6) sulfur (SINT), or (7) phosphorus (PINT), respectively
- (8) FULLP – the control group of the phosphorus deficiency experiments. Experiments for phosphorus deficiency were done separately from nitrogen and sulfur ones and our industry partners wanted to be able to identify the two control groups with 2 distinct labels (FULL and FULLP) in case of bias.

Data are generated via spectroscopy of leaves, using a VIS-NIR 350-2500 nm spectrometer, in other words from near ultraviolet to infrared. Leaves come in three types (indicated by the attribute L):

- $L1$: are leaves that sprout early and grow mainly in an environment where all nutrients are available. It is *expected* that they can store everything necessary and they are not expected to show deficiencies. However, spectroscopic measurements are done *after* the beginning of the nutrient deficiency.
- $L3$: sprout when all nutrients are available but potentially grow in a nutrient-deficient environment.
- $L5$: sprout already in a nutrient-deficient environment.

TABLE I
DIFFERENT NUTRIENT CONDITIONS AND THEIR PROPORTION OF THE DATA

Condition	Percentage of data
FULL	9.08%
N0	10.9%
S0	11.3%
P0	15.7%
NINT	10.9%
SINT	10.9%
PINT	15.7%
FULLP	15.5%

Our industry partners recommended that we ignore wavelengths $< 450nm$ and $> 2000nm$, which still left us with data of a dimensionality of 1551. Wavelengths outside this range have a too high noise-to-signal ratio to be useful. In addition to the wavelengths, data are described by the type of leaf, and the day (since the transplantation into the nutrient solution) when the measurement was taken, with possible values $\{9, 13, 16, 20, 23, 27, 30, 34, 37\}$. That attribute is denoted by *DaT* (Day After Transplantation). It should be noted that the deficient solutions are introduced on $DaT = 8$. All in all, we were supplied with 1365 leaf samples, with different nutrient conditions having the proportions shown in Table I.

Both our industry partners and we feel that using data analysis for optimizing plant nutrition is a promising future research direction, which is why we make the data available at https://figshare.com/articles/cleaned_spectra_csv/12445574.

V. A preliminary study of the spectroscopic measurements

In this section, we perform an initial analysis of the spectroscopic measurements as usual in a data science process to get a better understanding of the data. We find that the measurements differ according to the nutrient deficiencies so using these data to characterize the nutrient deficiencies is justified, but also that they are not simple to exploit.

a) Mean of wavelength measurements over all samples per condition.: To gain a first understanding of the data, and following the recommendation of our industry partners, we separated samples into different deficiency conditions and calculated the average value per wavelength. As Figure 1 shows, there are visually noticeable differences: plants experiencing phosphorus deficiency exhibit higher reflectance values for all wavelengths (note that FULLP and PINT are very close and hardly distinguishable). As for the other conditions, they are more or less well separated depending on the wavelength range. This indicates both that predicting deficiency conditions based on spectroscopy should be possible, and that it is not straight-forward.

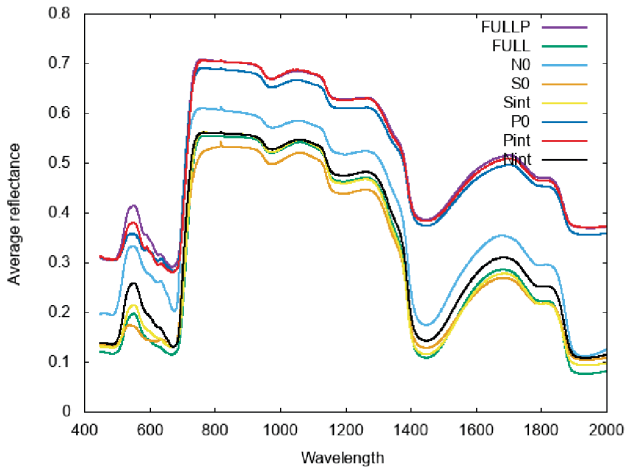


Fig. 1. Average wavelength reflectance according to wavelength per nutrient condition.

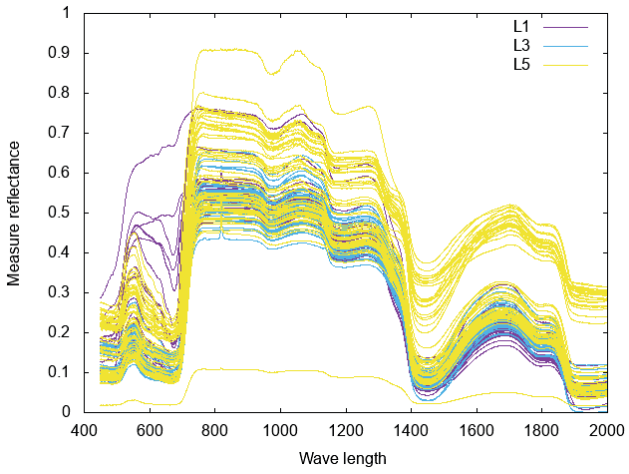


Fig. 2. For SINT, reflectance values according to wavelength ($L1$, $L3$ and $L5$ are leaf generations)

b) Reflectance values for all samples given a condition.: Figure 2 shows all samples belonging to the SINT class, color-coded depending on the leaf generation ($L1$, $L3$ and $L5$). As we can see, real data, especially one derived from a natural process, is not particularly homogeneous. Since this is already the case for these plants grown under controlled greenhouse conditions, we would expect that outdoor-grown plants to show even larger deviations. This problem that will be aggravated by a small-scale use of wide-spectrum spectrometers. This observation is also a motivation for the preprocessing techniques that we evaluate in Section VII.

VI. CHARACTERIZING DEFICIENCY CONDITIONS: DATE OF TRANSPLANTATION AND LEAF GENERATION

We start this section by defining a baseline in order to be able to evaluate preprocessing options objectively, then study the impact of the leaf generations and the date of transplantation in the context of characterizing deficiency conditions.

A. Establishing a baseline

The state of the art in predicting nutrient deficiency conditions applies discriminant analysis or regression methods. The most-widely used of these is partial least squares regression [10]–[12]. That is why we chose PLSR as a first baseline for classification. In concrete terms, we used the PLSR implementation included in the sklearn library, which requires binarized targets for classification. We therefore use one-hot encoding, representing each class label by an 8-bit vector, one entry of which is 1, the others 0. The output of a PLSR prediction being a numerical value, we interpret the output dimension closest to 1 as indicating the class to be predicted. We evaluated predictive accuracy via repeated randomized stratified 90%/10% training/testing splits. Our code is available at https://github.com/agrodm/dsaa.

Since phosphorus experiments are different from the ones for nitrogen and sulfur, and they did not expect the $L1$ generation to show any deficiencies at all, our industry partners recommended that we focus only on $L3$ and $L5$ leaves under nitrogen/sulfur stress. In addition, to arrive at better separability, they suggested working only on complete deficiencies (N0, S0), in other words removing NINT and SINT samples. This left us with a rather small data set containing 233 samples we call SD3C (for Small Dataset 3 Classes, – FULL, N0, S0). Our experiments show that SD3C is very easy to model w.r.t this simplified classification task (first row of Table II), a surprising result for our industrial partners. By contrast, as the second row shows, the full data set makes for a more challenging problem (experiments with more than 20 runs gave similar results).

As mentioned in Section II, in [7] two stress conditions, nitrogen deficiency and weed presence, were predicted. They report that while combined stresses can be predicted with only 69% accuracy, for nitrogen deficiency alone 81% can be reached. They report best results for an SVM with the RBF kernel, we therefore also used Weka's [13] SMO implementation with the RBF kernel and explored $\gamma \in [2^{-15}, 2^{3}]$ and $C \in [2^{-5}, 2^{15}]$ via grid search. We evaluated predictive accuracy via a stratified ten-fold cross-validation. The results are reported in the bottom half of Table II and we see a similar phenomenon as for PLSR, with the small data set **much** easier to model than the full one. After discussion, we convinced our partners to drop their suggested setting for the sake of deriving additional insights, and we will work on the full data set (FD) going forward.

In a real-life setting, we cannot expect to know the leaf generation, however, and there is a risk that the SVM exploits this information to split deficient leaves from non-deficient ones. The second column of Table II shows the effect of leaving out the leaf generation, and we can see that for the full, non-preprocessed data, removing leaf information indeed slightly lowers predictive accuracy for the SVM but slightly improves it for PLSR.

TABLE II
PREDICTION ACCURACIES FOR PLSR AND SVM USING THE RAW DATA. "DIM." DENOTES THE NUMBER OF WAVELENGTHS IN THE DATA

Data set (Dim.)	w/L & w/*DaT*	w/o L	w/o L & w/o *DaT*
PLSR			
SD3C (1551)	$99.08\%(100runs)$	$99.95\%(100runs)$	N/A
FD (1551)	$72.48\%(20runs)$	$73.65\%(20runs)$	$75.44\%(20runs)$
SVM w/RBF kernel			
SD3C (1551)	99.57%	99.57%	99.57%
FD (1551)	79.85%	79.34%	77.44%

Day after transplantation (*DaT*) is obviously also not available in a real-life setting (a farmer *would* know when he brought out the seeds, though). The third column shows the results when neither *DaT* nor L are available and we see a clear reduction in accuracy for the SVM (but also another improvement for PLSR). Finally, we also see that the SVM achieves higher accuracies than the PLSR for all settings. We will therefore only report SVM results going forward.

We'd like to stress that the use of the SVM is mainly due to its having been used in the literature before. While other, more recent techniques, might give better classification results, we mainly aim to characterize the differences between different representations, i.e. relative accuracy differences rather than absolute ones.

B. Addressing question 1: the impact of the time since the onset of the deficiency, i.e. DaT

The first question we were interested in is from what day on we can reliably predict nutrient conditions. We consider this question from two perspectives. The first is to simply test for each day how well (or not) samples of this day can be classified. Figure 3 (a) shows the results, for the moment we are only interested in the curve labeled "Raw data"[1]. Unsurprisingly, samples of *DaT*$= 9$ are difficult to classify, given that deficiencies have been introduced only a day earlier. Accuracy quickly ramps up, though, reaching almost 90% for *DaT*$= 34$ (the decrease after *DaT*$= 34$ is likely explained by the fact that at this day leaves are so deficient from nutrients that spectroscopic measurements are no longer accurate). It should be noted, however, that this represents a situation where the model is trained on samples from that day, a situation that we will rarely achieve in reality since we cannot know the onset of nutrient deficiencies.

An arguably more realistic option is therefore to successively remove earlier samples, which will help us to understand from which day on samples become informative because they begin to show more intense effects of the deficiencies. Figure 3 (b) shows the results for this setting. Note that while in the left figure all samples are of a certain *DaT*, on the right samples are of that *DaT* or later. We see indeed that accuracy increases, albeit with a drop at *DaT*$= 23$, gaining about seven percentage points by *DaT*$= 30$, compared to the full data. The decrease after *DaT*$= 30$ is probably due to the fact that only few samples of mainly highly nutrient-deficient leaves are left at that point. Finally, we see that already from *DaT*$= 13$, i.e. less than a week after the onset of nutrient deficiencies, rather good prediction is possible, these samples are therefore already informative.

[1]We will return to the other curves in Sections VII-B and VII-C.

C. Addressing question 2: the impact of the leaf generations

The second research question is whether all three leaf generations are useful for predicting nutrient deficiencies, and if so to what degree. Since, according to the experimental setup, different leaf types have undergone different deficiency conditions, there should be a relationship between leaf generations and nutrient deficiencies.

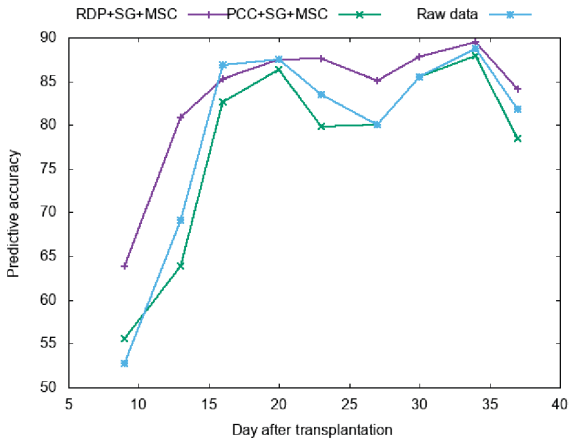
To answer this question, we look at the dual problem: given a *nutrient deficiency*, can we reliably predict *leaf generations*? We report classification results in Table III: for each of the six deficiency conditions, we trained an SVM with an RBF kernel to predict whether samples fall into $L1$, $L3$, or $L5$. Accuracies were derived via a stratified ten-fold cross-validation. SG, MSC are preprocessing techniques that we will discuss in Section VII-A, RDP and PCC dimensionality reduction techniques discussed in Sections VII-B and VII-C.

As the results in Table III show, this is indeed possible with rather high accuracy, even if the SVM never reaches 100%. For $L3$ and $L5$ leaves, this is partly an expected result, even though the high accuracy is encouraging, but as we have described above, $L1$ leaves should be able to store enough nutrients to not be affected by deficiencies, according to our partners. It is therefore possible that $L1$ leaves are simply so different from later ones that they are easy to qualify.

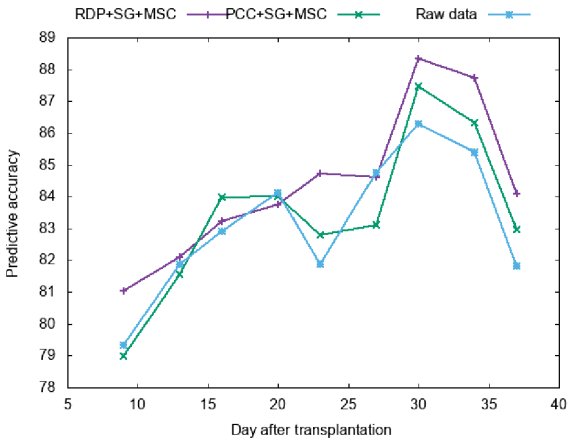
To investigate this further, we looked in greater depth into the role of $L1$ leaves. As Table IV shows, if we limit the data to $L1$ leaves only, we can still achieve predictive accuracies that are significantly higher than chance. The largest class contains only 17% of the $L1$ samples, whereas the worst result is at 65%. It seems therefore as if even $L1$ leaves experience the effects of nutrient deficiencies, which should allow for an earlier detection, which is an interesting result from the application point of view (and might partially explain our results for *DaT*).

VII. CHARACTERIZING DEFICIENCY CONDITIONS: IMPROVING THE DATA REPRESENTATION AND HIGHLIGHTHING USEFUL WAVELENGTHS

In this section we discuss the main part of our work: how to improve the data representation. We evaluate both direct options – smoothing and filtering – in Section VII-A, and



(a) Predictive accuracies according to the date of transplantation for different data representations.



(b) Predictive accuracies for successively smaller *DaT* intervals (i.e. samples with *DaT* or later) for different data representations.

Fig. 3. Effect of time since transplantation on predictive accuracies

TABLE III
PREDICTIVE ACCURACIES FOR PREDICTING LEAF GENERATION FOR DIFFERENT DEFICIENCY CONDITIONS

Data	N0	NINT	S0	SINT	P0	PINT
FD	91.28%	93.96%	92.21%	88.59%	92.52%	91.59%
FD+RDP+SG+MSC	92.62%	91.28%	90.91%	85.91%	91.59%	87.85%
FD+PCC+SG+MSC	91.28%	87.25%	90.91%	81.21%	91.59%	83.65%

TABLE IV
PREDICTIVE ACCURACIES USING ONLY $L1$ LEAF INFORMATION

Data	Accuracy
FD	68.51%
FD+RDP+SG+MSC	68.75%
FD+PCC+SG+MSC	65.39%

indirect ones – dimensionality reduction – in Sections VII-B and VII-C. The latter part addresses research question 3.

A. Filtering and smoothing the data

As shown in Section V, raw spectroscopic data are not homogeneous. One way of addressing the variability of spectrographic measurements is to filter and smoothen the data. To do this, we applied a Savitzky-Golay (SG) filter[2] [14], followed by multiplicative signal correction (MSC)[3] [15], both common preprocessing steps in spectroscopy.

As Table V shows, applying SG does not improve classification accuracy for the full data compared to Table II but the resulting representation weathers the removal of leaf and *DaT* information a bit better. Additionally applying the MSC algorithm, however, boosts accuracy by between one and two percentage points.

[2]Available in the `scipy.signal` library.

[3]Which we implemented ourselves. Our code is available at https://github.com/agrodm/dsaa

B. Dimensionality reduction using Ramer-Douglas-Peucker

Even though using SG and MSC *does* increase classification accuracy, these results were derived from the full spectrums of wavelengths. Yet, as stated before, one of our goals is dimensionality reduction to make data acquisition easier. An existing option for dimensionality reduction is to use the Ramer-Douglas-Peucker algorithm (RDP)[4], a method that reduces the number of points describing a curve by keeping only the start and end points of line segments [16]. We ran RDP with $\epsilon = 0.001$.

As Table VI shows, applying RDP improves predictive accuracy over the baseline (cf. Table II) and, more interestingly, creates a representation where removing leaf information leads to *additional* improvement.

Both the improvement and the significant reduction in the number of wavelengths are promising, and an obvious follow-up question is whether applying filtering and smoothing to this reduced representation allows for additional gains. Table VI also shows classification accuracies for when SG and MSC have been applied. Applying SG alone does not have an effect but additionally using MSC boosts accuracy on the full data by another $\sim$ 1.5 percentage points. Interestingly enough, removing leaf information reduces accuracy below that of the other two representations, yet when we also remove the *DaT* feature, we get the best result for that column.

[4]Available at https://pypi.org/project/rdp/

TABLE V
PREDICTION ACCURACIES USING AN SVM WITH RBF KERNEL, USING SMOOTHED AND FILTERED DATA. "DIM." DENOTES THE NUMBER OF WAVELENGTHS IN THE DATA

Data set (Dim.)	w/L & w/*DaT*	w/o L	w/o L & w/o *DaT*
FD+SG (1551)	79.85%	79.41%	77.51%
FD+SG+MSC (1551)	80.88%	80.29%	79.78%

TABLE VI
PREDICTIVE ACCURACIES OF THE SVM, USING DIMENSIONALITY-REDUCED DATA

Data set (Dim.)	w/L & w/*DaT*	w/o L	w/o L & w/o *DaT*
FD+RDP (540)	80.73%	81.1%	79.56%
FD+RDP+SG (540)	80.73%	81.1%	79.56%
FD+RDP+SG+MSC (540)	82.13%	81.03%	80.73%

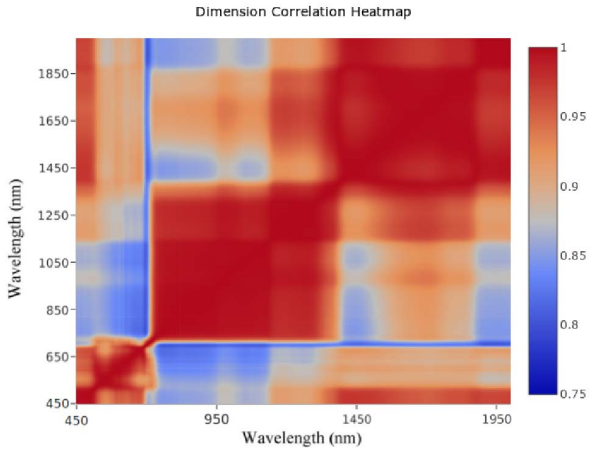


Fig. 4. Correlation between different wavelengths.

Going back to Figure 3, we see that the combined application of RDP, SG, and MSC outperforms learning and predicting on the raw data for almost all values of *DaT*. When it comes to predicting leaf generations, however, Table III shows that using the raw data is mostly preferable.

C. Dimensionality reduction via Pearson's correlation coefficient

The previous section has shown that RDP provides significant dimensionality reduction and, combined with SG and MSC, leads to good results on the data representation making the fewest assumptions. However, 540 separate wavelengths, though far less than the 1551 in the original data, still make for a quite a bit of spectroscopy. In this section, we therefore explore a different approach towards dimensionality reduction.

Inspired by RDP, we calculated the Pearson's correlation coefficient (PCC) between all pairs of wavelengths, resulting in the similarity matrix shown in Figure 4. Where RDP groups wavelengths according to the line segment they belong to, PCC measures how well pairs of wavelengths increase (or decrease) together.

The redundancy between adjacent wavelengths is clearly visible – in fact, as the legend on the right-hand side shows, the correlation coefficient is always ≥ 0.75. Yet we can also see split points where this redundancy is much lower than for adjacent areas. A straight-forward approach to dimensionality reduction groups wavelengths together for which the correlation coefficient exceeds a certain value. By setting the entries larger than a threshold in the similarity matrix to 1, 0 otherwise, we can visualize such groupings. Figure 5 shows them for four thresholds.

We want groupings that are rather homogeneous, well-separated, and help with dimensionality reduction. While a threshold of 0.999 does not help much with the latter, the choice between the other three thresholds was more of a judgment call. In the end, 0.95 grouped too many wavelengths together, especially the two thin bands at the left/bottom, whereas 0.98 already hints at lower correlations that fully detach for 0.99, the threshold we chose. Using dynamic programming, we can identify maximally sized sub-matrices containing only 1s (Figure 6, left), each of which can be represented by a single wavelength, achieving strong dimensionality reduction while not losing too much information. We retained only submatrices that group at least fifty wavelengths to smooth effects specific to the data sample.

A final important consideration is to avoid grouping wavelengths of different colors:

- 450nm - 500nm for blue
- 500nm - 600nm for green
- 600nm - 700nm for red
- 700nm - 1400nm for near infrared (NIR)
- 1400nm - 2000nm for shortwave infrared (SWIR)

This is particular important since different regions correspond to different phenomena: visible wavelengths represent leaf pigmentation, NIR cellular structure, and SWIR water retention. Taking this constraint into account narrows the groupings somewhat as can be seen in Figure 6, right. As we can see on the left-hand side, however, even unconstrained groupings recover known regions rather well.

Together, the seven groupings cover 1295 wavelengths, meaning that we reduce dimensionality from 1550 to 255, a greater reduction than using RDP.

Table VII shows prediction accuracies using the PCC dimensionality-reduced data, on raw, filtered, and filtered and

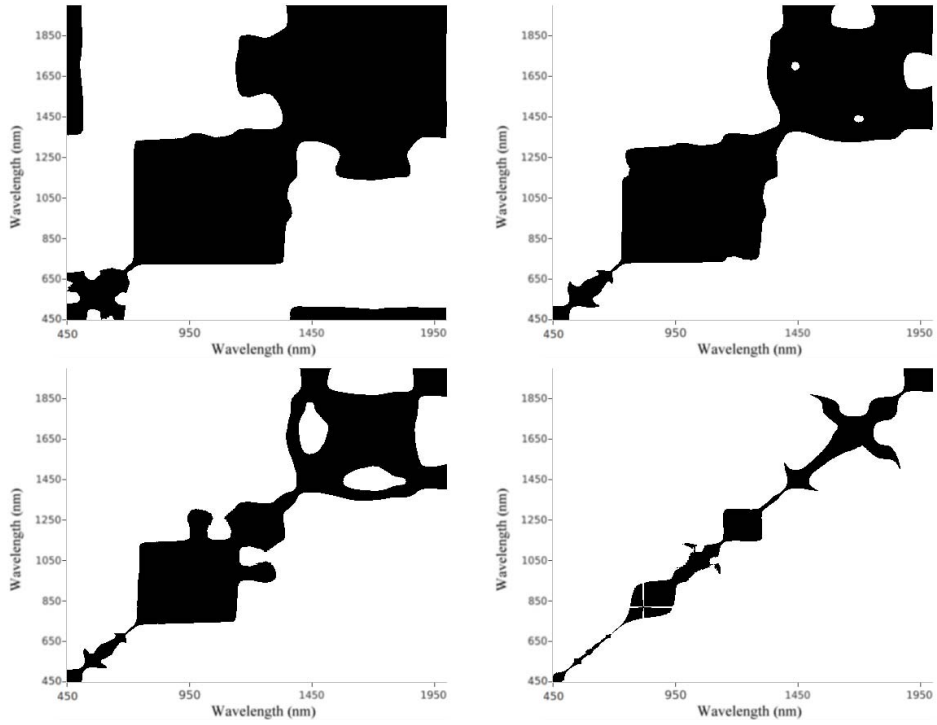


Fig. 5. Binarized correlation matrices, for cutoffs 0.95, 0.98 (top row), 0.99, and 0.999 (bottom row).

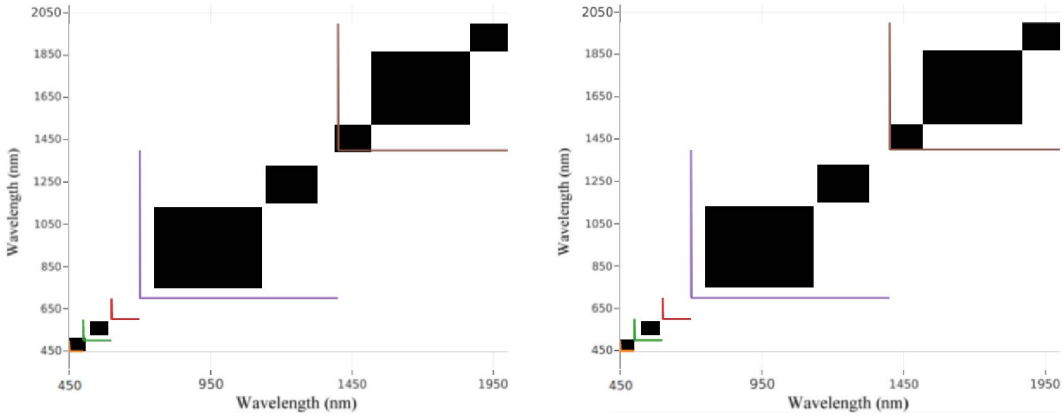


Fig. 6. Maximally sized squared matrices of strongly correlated wavelengths (left-hand side), taking color constraints into account (right-hand side).

TABLE VII
PREDICTION ACCURACIES USING AN SVM WITH RBF KERNEL, USING SMOOTHED AND FILTERED DATA

Data set (Dim.)	w/L & w/DaT	w/o L	w/o L & w/o DaT
FD+PCC (255)	78.02%	77.44%	75.09%
FD+PCC+SG (255)	77.95%	77.44%	75.24%
FD+PCC+SG+MSC (255)	79.78%	78.97%	75.53%

smoothed data. As before, the main improvement to accuracy comes from the MSC correction, even though SG also slightly improves accuracy for the setting without *DaT*. In comparison to Table VI, accuracies are lower, especially for the setting with *DaT*. As Figure 3 shows, the representation PCC, SG, and MSC is outperformed for individual *DaT* but mostly performs better for the ablation setting. The most interesting result of Table III is that the representation shows a relatively big drop-off for leaves showing partial deficiencies.

All in all, representations derived using PCC for feature reduction perform worse than those using RDP. Yet PCC-reduced data retain less than half of the wavelengths for RDP-reduced one. This would allow for less expensive spectroscopy, and therefore more and more precise measurements.

A word of caution: all the dimensionality reduction methods in this work assume directly or indirectly that there is a semantic connection between adjacent features. In fact, they all exploit the fact that significant changes between adjacent features mean that the feature should be kept. While they are therefore well-suited to the kind of spectrographic data that we have explored in this work, different kinds of descriptors might benefit less from the proposed method.

VIII. Conclusion and perspectives

Detecting nutrient deficiencies is a major challenge of sustainable agriculture since it allows to improve soil quality by bringing plant nutrients as much as necessary but as little as possible. In this paper, we have discussed experiences with a real-life agronomic data set. Our industry partners approached us with certain preconceived notions about their spectroscopy data, and a concrete problem setting: predicting nutrient deficiencies during plant development, ideally quickly, reliably, and at not too high cost.

So far, this problem setting is underexplored in the literature, and we therefore attacked it from different angles. We first looked at whether it is in fact possible to predict deficiencies early, and find that (at least under controlled conditions) high-quality predictions are possible less than a week after onset of the deficiency. A second question was to what degree deficiencies are expressed in the plant's leaves. According to our partners' intuition, leaves that mainly grew in a nutrient-rich environment should be useless in terms of identifying recently arrived deficiencies, pushing back the detection time. As we could show empirically, however, this intuition is somewhat incorrect, meaning that identification and correction should be possible much earlier than assumed. This result underlines the bias that can be introduced in spite of themselves by the experts on the data and we think that this remark can be generalized to many other problems.

Finally, we explored options for tweaking the data to improve classification accuracy and/or reduce the cost of future data acquisition. By exploring two dimensionality reduction techniques (RDP and PCC), we proposed a new data representation that allows to improve on the results on the raw data while reducing the number of wavelengths used to describe the data by two thirds. The latter, while losing somewhat in terms of predictive accuracy, halves the number of wavelengths RDP yet again, which could allow a much cheaper data acquisition process, allowing the analysis of more samples and more precise measurements.

While we have shown how to apply different preprocessing techniques to this kind of data, we need to go further. An important bottleneck is simply data: after all, we were in the fortunate situation that the industry partners we collaborated with have their own experimental greenhouse but even with this situation, we had only access to $\approx$1300 data points for the spectral data. As a next step we will design further greenhouse experiments in collaboration with our partners, analyzing samples based on the results of the work described in this paper, i.e. measuring smaller ranges of wavelengths.

An additional direction we explored, but do not report on the paper, is that of subgroup discovery. Subgroup discovery is the task of identifying descriptions that describe subsets of the data in which a value of a target variable is over-represented in comparison to the full data. Finding combinations of wavelengths whose values are characteristic of certain deficiencies could allow the construction of specialized sensor setups that involve only a small range of wavelengths, and in addition offer interpretable patterns that could help in understanding the effects of deficiencies in more detail. Unfortunately, straightforward application of an off-the-shelf subgroup discovery system, while producing intriguing patterns, did not yield satisfactory results yet. Between descriptions that are not specific enough and too much redundancy in the result set, we could not properly characterize deficiency conditions. We intend to explore this direction with a more tailor-made solution going forward.

References

[1] K. Liakos, P. Busato, D. Moshou, S. Pearson, and D. Bochtis, "Machine learning in agriculture: A review," *Sensors*, vol. 18, no. 8, p. 2674, 2018.

[2] S. Dimitriadis and C. Goumopoulos, "Applying machine learning to extract new knowledge in precision agriculture applications," in *Panhellenic Conf. on Inf.* IEEE, 2008, pp. 100–104.

[3] R. O. Kuchenbuch and U. Buczko, "Re-visiting potassium-and phosphate-fertilizer responses in field experiments and soil-test interpretations by means of data mining," *J. of Plant Nutrition and Soil Science*, vol. 174, no. 2, pp. 171–185, 2011.

[4] H. Yu, D. Liu, G. Chen, B. Wan, S. Wang, and B. Yang, "A neural network ensemble method for precision fertilization modeling," *Math. and Comp. Modelling*, vol. 51, no. 11-12, pp. 1375–1382, 2010.

[5] A. Shekoofa, Y. Emam, N. Shekoufa, M. Ebrahimi, and E. Ebrahimie, "Determining the most important physiological and agronomic traits contributing to maize grain yield through machine learning algorithms: a new avenue in intelligent agriculture," *PloS One*, vol. 9, no. 5, 2014.

[6] J. Yuan, C. Liu, Y. Li, Q. Zeng, and X. F. Zha, "Gaussian processes based bivariate control parameters optimization of variable-rate granular fertilizer applicator," *Computers and Electronics in Agriculture*, vol. 70, no. 1, pp. 33–41, 2010.

[7] Y. Karimi, S. Prasher, R. Patel, and S. Kim, "Application of support vector machine technology for weed and nitrogen stress detection in corn," *Computers and electronics in agriculture*, vol. 51, no. 1-2, pp. 99–109, 2006.

[8] J. Gholap, A. Ingole, J. Gohil, S. Gargade, and V. Attar, "Soil data analysis using classification techniques and soil attribute prediction," *arXiv preprint arXiv:1206.1557*, 2012.

[9] S. Jay, F. Maupas, R. Bendoula, and N. Gorretta, "Retrieving lai, chlorophyll and nitrogen contents in sugar beet crops from multi-angular optical remote sensing: Comparison of vegetation indices and prosail

inversion for field phenotyping," *Field Crops Research*, vol. 210, pp. 33–46, 2017.
[10] D. M. Haaland and E. V. Thomas, "Partial least-squares methods for spectral analyses. 1. relation to other quantitative calibration methods and the extraction of qualitative information," *Analytical chemistry*, vol. 60, no. 11, pp. 1193–1202, 1988.
[11] J. Li, W. Huang, L. Chen, S. Fan, B. Zhang, Z. Guo, and C. Zhao, "Variable selection in visible and near-infrared spectral analysis for noninvasive determination of soluble solids content of 'ya'pear," *Food Analytical Methods*, vol. 7, no. 9, pp. 1891–1902, 2014.
[12] P.-T. Guo, Z. Shi, M.-F. Li, W. Luo, and Z.-Z. Cha, "A robust method to estimate foliar phosphorus of rubber trees with hyperspectral reflectance," *Industrial Crops and Products*, vol. 126, pp. 1–12, 2018.
[13] E. Frank and I. H. Witten, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, 1999.
[14] A. Savitzky and M. J. Golay, "Smoothing and differentiation of data by simplified least squares procedures." *Analytical chemistry*, vol. 36, no. 8, pp. 1627–1639, 1964.
[15] H. Martens and E. Stark, "Extended multiplicative signal correction and spectral interference subtraction: new preprocessing methods for near infrared spectroscopy," *Journal of pharmaceutical and biomedical analysis*, vol. 9, no. 8, pp. 625–635, 1991.
[16] D. H. Douglas and T. K. Peucker, "Algorithms for the reduction of the number of points required to represent a digitized line or its caricature," *Cartographica: internat. journal for geographic information and geovisualization*, vol. 10, no. 2, pp. 112–122, 1973.