Leiden University
Faculty of Science
Leiden Institute of Advanced Computer Science
Computer Science Curriculum

Bob van Schendel

Qualitative data from quantitative seed oxygen measurements

Supervisors: Dr. M.T.M. Emmerich
Dr. L. Cao
Prof.dr. A. van Duijn

Leiden 2021

Abstract

Horticulture is an area of human knowledge that has been in development since the literal human prehistory. Trial and error and empirically derived knowledge allowed us to thrive in almost any area on Earth. In modern day society horticulture has moved out of the public eye as food production has become very efficient, and as a result is now the domain of very few, very specialized people. With most of the population in developed countries not being involved in food production at all. Despite this, horticultural development and research has only accelerated. Processes to probe, monitor and develop plants have become a real science with ample supply of data about seeds. Optimizing plants for the production of food and medicine is the current goal. A recently developed method of assessing seeds is by directly monitoring their respiration. This respiration data can then be used to characterize the behavior of the seeds in question and provide us with an understanding of their vigor. We will show that we can use these characteristics to reliably predict their germination rate, as well as identify interesting statistical relations. This is encapsulated in an easy-to-use tool, in order to provide anyone in the horticultural industry access to this kind of analysis about seeds of their interest.

Contents

1 Introduction

1.1 Background

In horticulture, the specific species and type of seed used for planting depends greatly on the climate, soil and other characteristics of the local ecosystem. For a given harvesting goal, knowing the characteristics and quality of seed is very important as this is essential for maximizing harvesting yield. The characteristics that make the seed thrive in given conditions are known as the **vigor**. These characteristics include, for example, the chance of successful germination or the resistance to cold or drought. Ideally, given a batch of seeds, we can test these characteristics to know where and when the batch can be maximally utilized. In practice, this is not quite so simple. One way of assessing the vigor of a seed batch, we should simply plant them in the intended soil under the intended conditions and monitor the germination. This can be done at a small scale with a representative sample of the batch. This sample would be planted, watered and kept under supervision to identify radicle emergence of the seed to then infer the vigor of the batch statistically. Radicle emergence means the exact moment that the husk of the seed breaks and the root of the seed comes out. Unfortunately, this method is very slow and labor-intensive, as each individual seed must be under constant supervision in order to identify germination.

An alternative to this labor-intensive method of manual seed tracking is to record the phenomenological characteristics with a machine. These properties are then used to estimate the vigor of the seeds. Certain statistical relations have been established in the past, such as the relation between seed respiration and germination [4]. This respiration can be recorded by, for example, the SRA[1] which is developed by the company *Fytagoras*[2]. This machine records the respiration and outputs this data in a format that can then be automatically analyzed by computer programs. This approach eliminates the need for manual supervision of the seeds and, given sufficiently accurate statistical approaches, is reliable in estimating the actual radicle emergence of given seeds.

Various assays have been described to estimate seed vigor in the literature such as measuring the electrical conductivity or recording the dehydrogenase activity [2]. These methods, however, take time and are less practical than directly measuring the oxygen respiration during germination.

Although oxygen respiration and total germination have been directly linked before [1], there are few specific results. Furthermore, these results only indicate a positive correlation between well-shaped oxygen consumption curves and odds of germination. For further respiration analysis we need more knowledge on what we can infer from specific characteristics of these curves.

[1] Seed Respiration Analyzer

[2] https://www.fytagoras.com/nl/

1.2 Seed Respiration

1.2.1 Seed Respiration model

The model we use for seed respiration has the following assumptions:

1. Each seed is enclosed in an airtight capsule.
2. Each seed has a limited amount of air to use for respiration in the capsule.
3. Each seed is ungerminated before being entered into the capsule.
4. During germination each seed will only consume oxygen as they have not yet formed any chlorophyll. As a result the oxygen level will drop monotonically.

When these seeds are put inside the capsule they slowly increase their respiration to begin the germination process. After the oxygen level inside the capsule drops to the level where the seed cannot effectively respire anymore the respiration slows and the seed slowly stops its germination process. This process leads to a distinctive curve, reminiscent of a sigmoid curve, see Figure 1 for an example of this likeness.

Our model takes into account earlier work by the IIRB work group 'Seed Quality and Testing' [3] by using the so-called *ASTEC* values. These are values that represent characteristics of oxygen curves and are considered indicators of a seed's vigor. As the seed begins to repair the molecular germination mechanisms it slowly increases its oxygen consumption. After this repair is mostly completed the oxygen consumption of the seed is greatly increased and we call this moment the **Increased Metabolism Time** (**IMT**). The critical level of oxygen where seed respiration drops off relatively rapidly is called the **Critical Oxygen Pressure** (**COP**). The minimal value of the derivative is called the **Oxygen Metabolism Ratio** (**OMR**). The **IMT** and **COP** are used for the prediction of germination time by using a linear equation that crosses both these points. The x-intersect of this line is taken as the **Relative Germination Time** (**RGT**).

1.2.2 Seed Respiration Analyzer

The Seed Respiration Analyzer mentioned in Section 1.1 is a machine developed by *Fytagoras* that automates the tracking of oxygen level over a given timeframe. It can be loaded with a large number of seeds that are individually contained in capsules. It then automatically measures the oxygen level in each capsule over a specified period of time with a given interval between measurements. It outputs the recorded data as a file that can be analyzed by a computer program. This data then consists of the oxygen level at the timepoints of the measurements and this is categorised per seed.

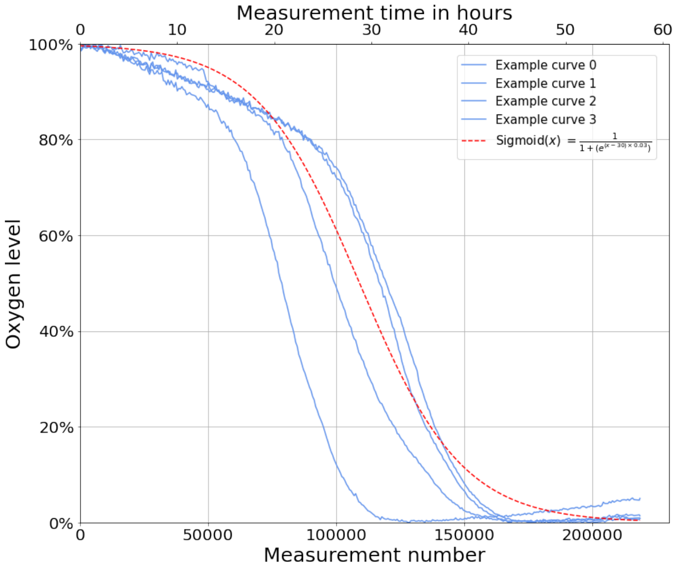


Figure 1. An example of 4 sigmoid-shaped oxygen curves over time showing the similarity to an ideal sigmoid curve.

1.3 Data and Acquisition

The data consists of two parts: the oxygen measurements of given seeds over time and the germination data of the same seeds. The latter data set is used to verify the accuracy of our trained model and is not necessary to produce for a batch of seeds we want to process.

The oxygen measurements are usually made with the SRA and contain the information for 2 trays containing 96 seeds each. The germination process is recorded over several days and consists of the oxygen level at certain timepoints. These measurements inevitably involve a small error that differs for every measurement, we can see this in Figure 2. This measurement noise must be ameliorated to allow us to obtain reliable measurements from the curves. We can also see that the oxygen measurements roughly follow a sigmoid curve, as the theory in 1.2 describes. Furthermore we will use the term *curves* for a given oxygen measurement time series. The data set containing the manually verified germination information for a given batch is used to train our predictive model, in order to establish the reliability of the approach. This is simply a file containing seed names and a boolean value indicating whether that seed germinated or not at any point in the recorded time.

1.4 Research question

We will show that we can reliably extract descriptive statistics from the oxygen-over-time curves of the SRA. These statistics are used to expand on the *ASTEC* values used in earlier experiments [3]. These can be used to predict germination percentage to a high degree and provide an accurate picture of the vigor of seeds in a given batch. This process is automated

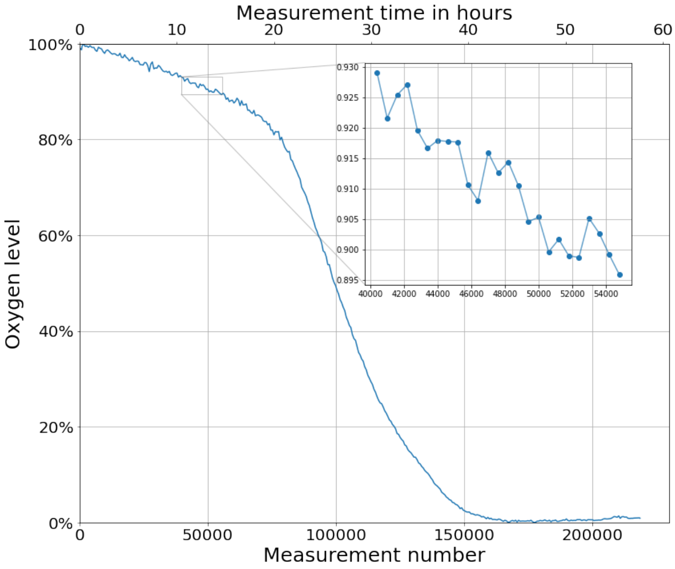


Figure 2. The Oxygen level over time for a certain seed.

and provided by means of a Jupyter notebook in a Colaboratory file and was written in Python 3. This notebook can be easily reconfigured and adapted to new statistical analyses, tests or visualizations.

More concretely the research questions are as follows:

1. What type of tool is necessary for the aid in analysis of oxygen data from seed germination?
2. How can this tool be used to analyze this oxygen data with minimal programming proficiency?
3. Which parameters and characteristics are useful indicators for the prediction of germination chance through the analysis of oxygen respiration?

The tool is detailed in the Section 5. The methods of answering these questions will be described in Section 2 and answered in Section 3.

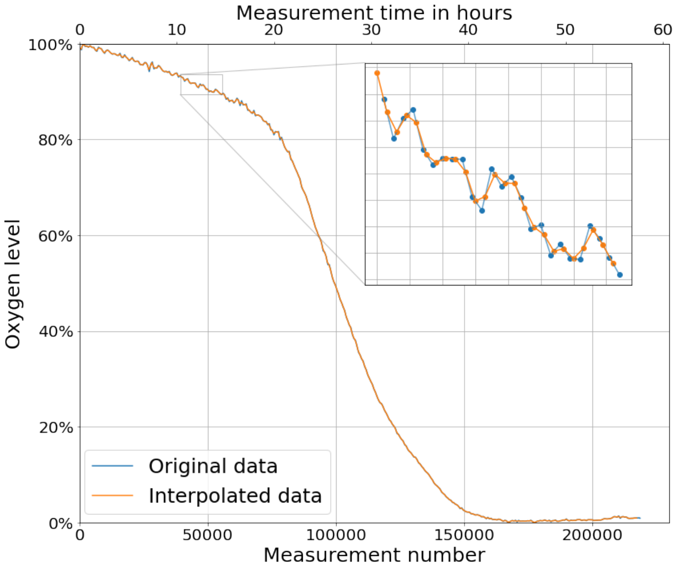


Figure 3. Linear interpolation applied to an oxygen curve.

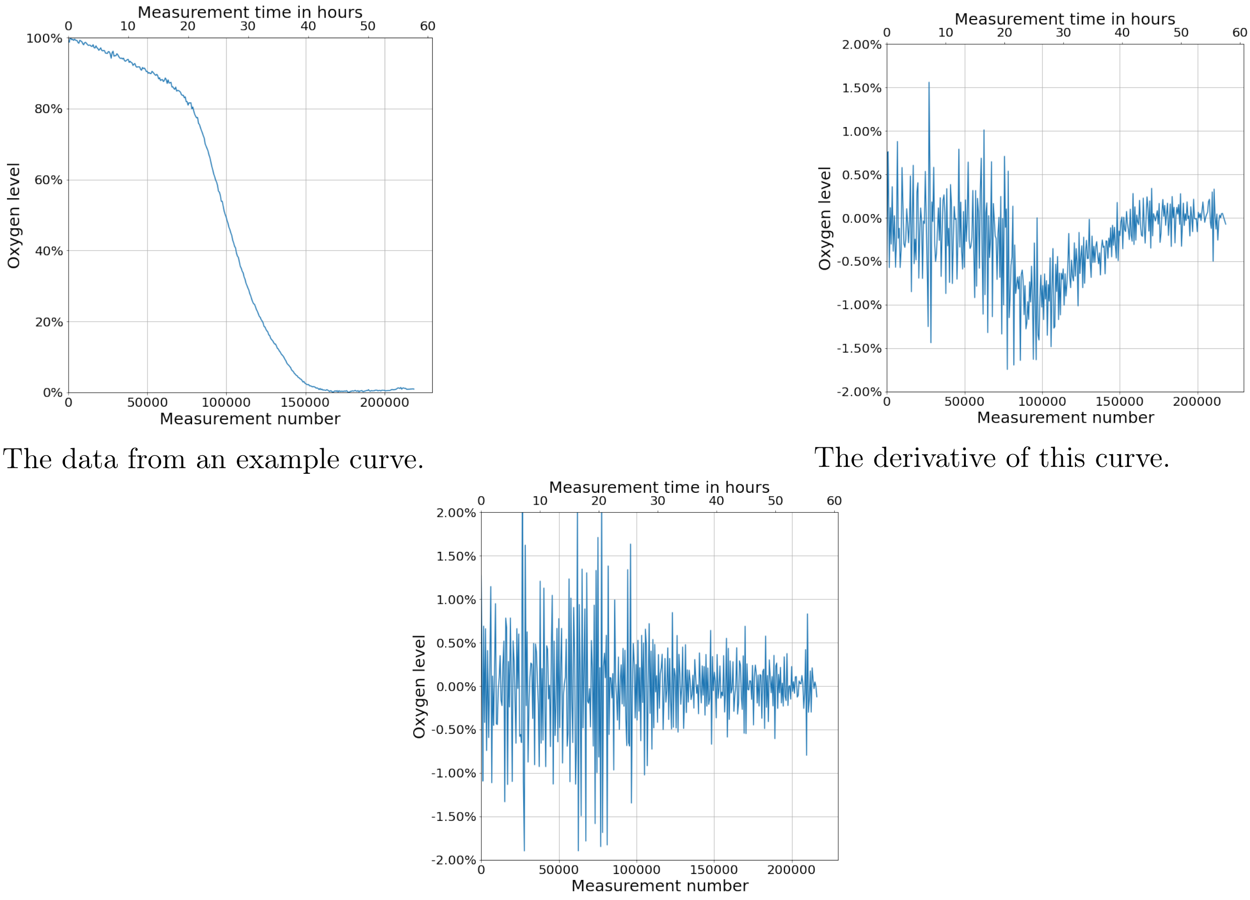
2 Methods

2.1 Preprocessing

As mentioned in Section 1.3 we have input data that consists of a csv file with the oxygen concentration of every capsule for each timepoint as well as some information about the testing that we won't use in our program. We reformat the data to produce a csv file of shape $M \times N$ where M is the number of measurements and N is the number of seeds. Each element is a float that represents the oxygen concentration of a specific seed at a specific time. This data then represents the raw oxygen level data from the curve. The distances between timepoints of the measurements are not completely uniform. Generally there are 600 seconds between the measurements, but this varies somewhat, mostly by only a few seconds. We apply linear interpolation to get uniform timepoints in the data. This is the method of choice as it allows us to get uniform times while not applying too much transformation, which could perturb our data for the actual analysis. An example is in Figure 3.

2.2 Curve Model Selection

The oxygen data is noisy so filtering must be applied before we can use it properly. Especially the first and second derivative, which we use to calculate characteristics of our curves, are not usable when calculated from the data directly. This can be seen in Figure 4. There are a number of ways we can apply noise reduction to our curves, most of them involve fitting a curve to our data.



The data from an example curve.

The derivative of this curve.

The second derivative of this curve.

Figure 4. An example curve to show how a little noise in the original data is amplified in the derivatives.

The goal of this noise reduction is to remove the noise from our curves while not removing the fluctuations that are present in them. After all, these represent information present in the curve. A way to assess the quality of the noise reduction is to see how well the resulting curve fits to the original line, while not having a positive derivative anywhere. This is in line with assumption 5 from Section 1.2.1. The options experimented with were:

1. Applying gaussian (or other) filters
2. Rolling means
3. Fitting sigmoid curves
4. Fitting a high-degree polynomial function
5. Fitting B-splines[3]

[3] B-splines are piece-wise polynomial functions that constrain these functions to be continuous across a number of their derivatives across each knot.

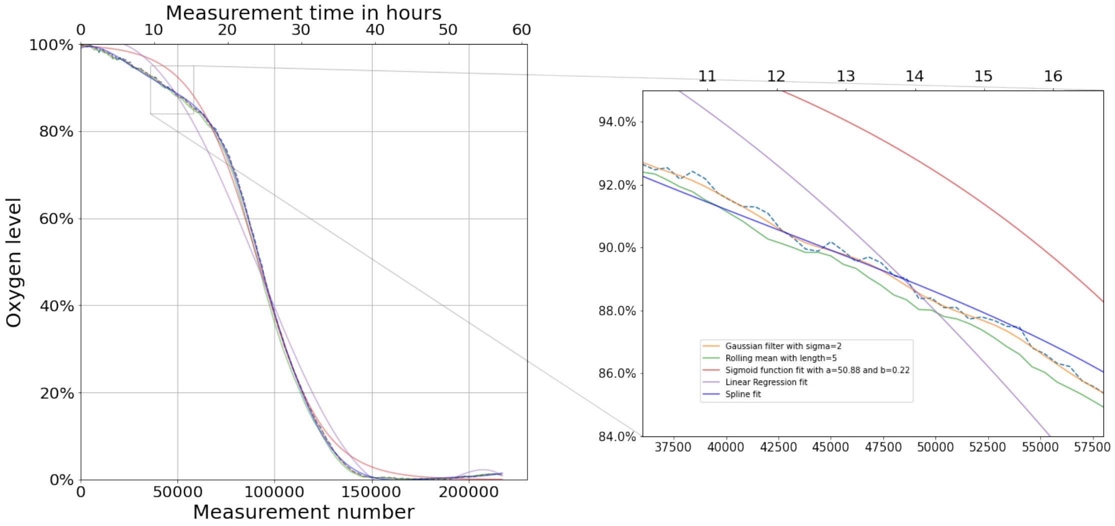


Figure 5. The interpolated data from a curve as well as 5 different models fitted to the data.

Fitting B-splines was chosen as an option for noise reduction. This method filtered out erroneous noise to a degree that allowed proper use of the derivatives calculated from the data, while also preserving the fluctuations in the oxygen level that might represent usable information. These splines have degree 4 and have 15 knots over the entire length of the curve. An example of all applied methods, as well as the chosen spline application is in Figure 5. These B-splines were then used to calculate the characteristics of the individual seed curves.

2.3 Descriptive statistics

When we have a curve fitted to the data, we can begin calculating statistics from it. We calculate a large number of descriptive statistics that function as the characteristics of the curve. This way we can forgo using the curve itself and instead use the descriptive statistics to attempt to find relations between the curve shape and the seed vigor. For this reason we calculate an exhaustive number of values in addition to the *ASTEC* values defined in Section 1.2.1. These additional values are:

- Curve maximum value
- Curve minimum value
- Curve average value
- Average derivative value
- Minimum derivative value
- R75, the first time the oxygen reaches 75%

- R50, the first time the oxygen reaches 50%
- R25, the first time the oxygen reaches 25%
- Maximum value of the curve's integral
- Average value of the curve's integral
- I50, the first time the curve integral reaches 50% of its maximum
- I25, the first time the curve integral reaches 25% of its maximum.

2.4 Seed vigor-curve shape relation

The overarching goal of this process is to be able to quantitatively relate the shape of the oxygen curve of a germinating seed to its vigor. Mainly, vigor is about the odds of germination of a seed. To do this we will attempt to relate the descriptive statistics of a curve to the information of whether a seed has germinated or not. Using a data set consisting of oxygen curves and manually verified germination statistics we can attempt to train a model to find this relation.

For the model, we use logistic regression as our dependent variable is binary and we intend to identify a linear combination of the input features that predicts the germination result. As we use a decision rule that maps the output to 1(**True**) or 0(**False**) this regression model effectively becomes a classification model. However, we will continue referring to the model as Logistic Regression as it is technically more correct. We will use 2 variations. The first is a plain logistic regression model that we will train on a training subset of our data. The second is a regularized logistic regression model that uses $L1$ regularization. $L1$ regularization has a tendency to set variables to 0 if they have a minute impact on the prediction accuracy. This way, varying the regularization parameters allows us to identify the significantly predictive properties of the curve. Combining this with a correlation matrix will allow us to identify relations in and between the descriptive statistics. We will use a Spearman's correlation coefficient r_s as our correlation measure as our variables are not all normally distributed. This value tells us when variables are monotonically related. The formula for Spearman's correlation coefficient is

$$r_s = \rho_{rg_X, rg_Y} = \frac{cov(rg_X, rg_Y)}{\sigma_{rg_X}\sigma_{rg_Y}} \tag{1}$$

Where ρ is the Pearson correlation coefficient *applied to the rank variables*, $cov(rg_X, rg_Y)$ is the covariance of the rank variables and σ_{rg_X} and σ_{rg_Y} are the standard deviations of the rank variables.

Having identified relations in our data, we will use significance testing to check the statistical significance of our findings. This will also tell us about the generalizability of these relations.

2.5 Software

The aforementioned operations are written in a software tool called a Jupyter Notebook. This is a program that runs a Python script interactively in a browser window without having to use the command line. All parts of this software tool are written in Python and most of the processing is done with **Numpy**, **Pandas**, **Scipy** and **scikit-learn**, which are software libraries for Python. The Jupyter Notebook in question is uploaded to a Google Colaboratory file, which makes running the code even easier. It allows any user with access to open the link to the file in a browser, upload the data (which also protects its privacy) and run the program without installing anything. This methodology makes this program widely usable by any authorized user, even those without extensive programming experience.

3 Results

3.1 Parameters

For the model we used the Sklearn.linear_model *LogisticRegression* class to create a Logistic regression model with the following parameters for the regularized model: $C = 6$, *penalty* = l1, *random_ state* = 0. The parameters for the unregularized model: *penalty* = none, *random_ state* = 0. For the C parameter we use the value 6, which is calculated as optimal by testing all values from 0.1 to 10 with increments of 0.1 in this regularized model. This model tests each value by doing 10-fold cross validation, with the *KFold* class in Sklearn.model_selection. The Mean Accuracy of these 10 splits are averaged and that is used as an indicator for the C value tested.

3.2 Training and testing data

For the training and testing subsets of the data we use the complete dataset of Tataros seeds, consisting of 384 oxygen curves. For the training and testing splits we use the *train_ test_ split* class in Sklearn.model_selection. The *test size* is set to 0.25 so 25% of the dataset will be used as a testing set and the remaining 75% will be used for training. We also shuffle the training and testing data by setting the *shuffle* parameter in the aforementioned class to *True*.

3.3 Linear model results

Training the non-regularized and regularized logistic model on our data gives us the coefficients visible in Table 1. These coefficients are obtained after fitting the model on a training subset of the data. These coefficients tell us how much influence each characteristic has on the germination prediction result. In turn we can use this as a way of selecting curve characteristics that might represent an interesting biological property relating to the chance of germination.

The Mean Accuracy for the L_1-regularized Logistic regression model was 0.885.

The Mean Accuracy for the unregularized Logistic Regression model was 0.865.

These accuracies are calculated over the test set as this gives a reasonably reliable idea of the accuracy for a different data set.

Model	L_1-regularized Logistic Regression	Logistic Regression
Max value	0.35	10.16
Min value	4.91	28.12
Mean value	0	-10.99
Max derivative value	0	-21.42
Mean derivative value	-4.63	-6.18
R75	0	-8.41
R50	-1.96	-4.23
R25	-1.79	-1.68
IMT	0.97	0.39
OMR	-4.85	-6.18
OMR timepoint	1.11	1.41
COP	-0.04	-1.18
RGT timepoint	-1.87	-4.04
Max integral value	0	-11.01
Mean integral value	0	3.26
Integral R50	0	11.87
Integral R25	0	10.07

Table 1. The coefficients for the separate characteristics of the trained logistic models.

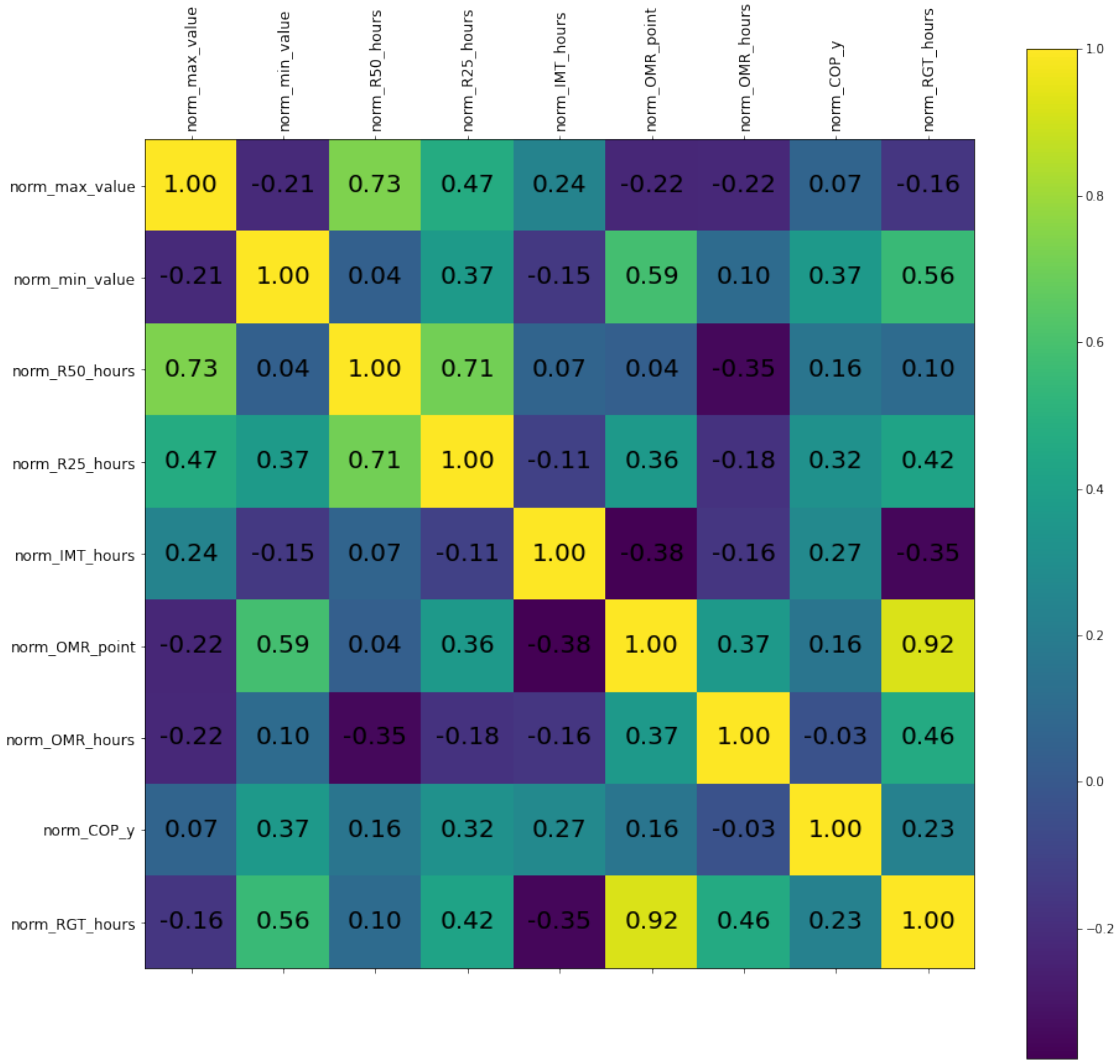


Figure 6. The Spearman's r_s correlation matrix calculated over the curve statistics that are found to be significant in the linear model.

3.4 Significant correlations

When we select these characteristics and plot the Spearman's r_s correlations of the entire dataset in a matrix we get the result we see in Figure 6. We will limit our analysis to the correlations ≥ 0.5 as correlations below this value are so weak they are not interesting for further analysis. These are the following:

1. Max curve value and R50 timepoint with $r_s = 0.73$.
2. Minimal curve value and OMR value with $r_s = 0.59$.
3. Minimal curve value and RGT timepoint with $r_s = 0.56$.
4. R50 timepoint and R25 timepoint with $r_s = 0.71$.
5. OMR value and RGT timepoint with $r_s = 0.92$.

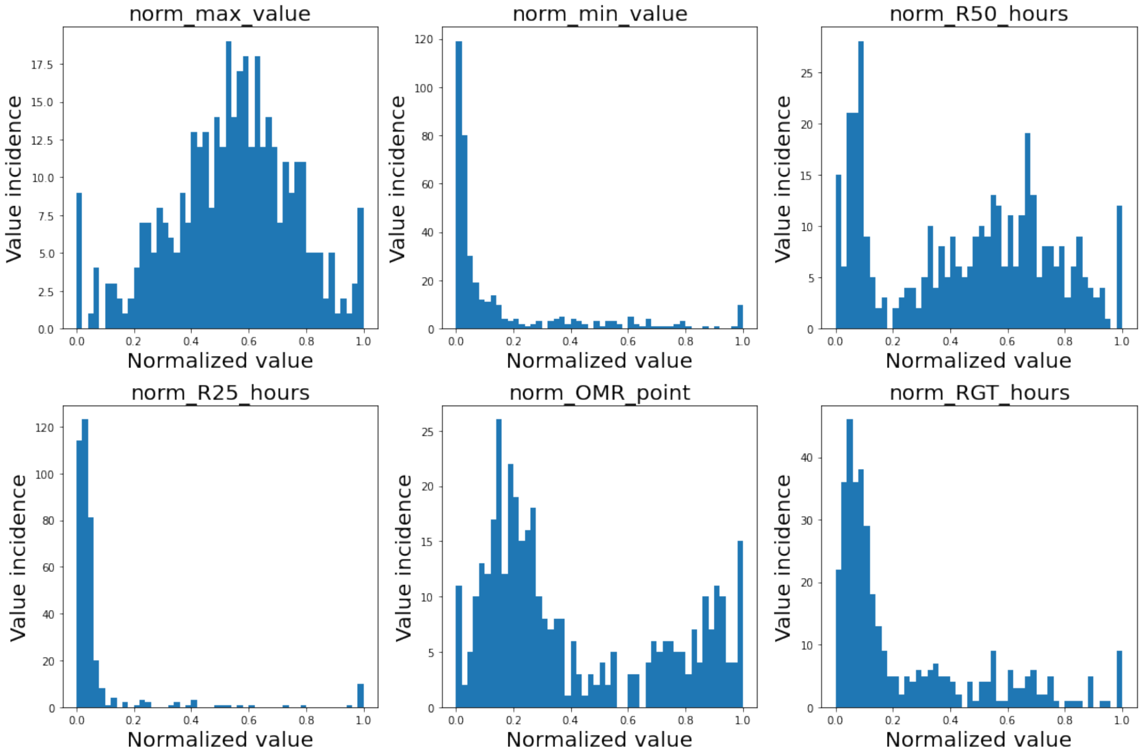


Figure 7. The Histograms of the variables found to have significant Spearman's r_s correlations.

We need to know if these variables are normally distributed and for that we plot the histograms of these values in Figure 7. We see that they are not normally distributed so to find out accurately what the correlation is and the significance thereof we will use *Spearman's* correlation coefficient instead of *Pearson's* correlation coefficient. We take our null hypothesis to be that there's no statistically significant correlation between these variables, so $r_s = 0$.

Our alternative hypothesis will be $r_s \neq 0$ and we will set our $\alpha = 0.05$. We calculate our test statistic for each correlation. T_1 refers to the first significant correlation found in Section 3.4 (between max curve value and R50 timepoint), T_2 refers to the second correlation, etc. We then calculate our test statistics.

$$\begin{align}
T_i &= \sqrt{\frac{n-2}{1-r_s i^2}} \tag{2}\\
T_1 &= 28.6 \tag{3}\\
T_2 &= 24.21 \tag{4}\\
T_3 &= 23.59 \tag{5}\\
T_4 &= 27.76 \tag{6}\\
T_5 &= 49.87 \tag{7}\\
& \tag{8}
\end{align}$$

We have $df = 384 - 2 = 382$ and find, with the help of a table of t-distribution values, that we have $P < 0.05$ for all correlations that we found. The test statistic is so far above the critical value that we can reject all the null hypotheses and conclude that we have found clear statistically significant relations in this data. Specifically, we have found that we have monotonic relations between the variables named in Section 3.4.

4 Conclusion

We have seen that a relatively simple tool intended for the analysis of oxygen data can make the analysis itself trivial. With the help of my program, a few lines of code can make a large body of data accessible and easy to analyze. This way the programmatical burden of analysis is removed from researchers who don't have the time or capability to develop such things themselves. And while tools like these help research, the horticultural industry also thrives by tools that speed up research and development work.

This tool was not only developed in an accessible language (Python), it is also fully usable and accessible in any modern browser. The code is encapsulated in a Jupyter notebook which is hosted on Google's Colaboratory website. This allows any user to go to this notebook via a link, upload data with drag and drop and run everything with a command that is as simple as pressing Control + F9.

The only development our analysis needed was to upload data, run the main program pipeline and then some simple *numpy* and *Pandas* commands to visualize our data and calculate things like statistics. We made a simple module containing the statistics we wanted to calculate about every seed and put it in place in the pipeline. These statistics could then be used to identify some statistical relations in the data (Section 3.4) and show that we could easily fit a simple model to reliably predict germination based on nothing but seed respiration data (Section 3.3).

5 Program documentation

The program expects files in the comma-separated value format. In files containing oxygen curves a column should hold the oxygen information for a single seed in a header at the top of the file. In files containing germination data every row should contain a seed name and corresponding boolean value and no header at the top of the file. Using the Colaboratory interface is the easiest way of uploading data to the environment of the notebook. After uploading the files can be referenced by using the filename and extension in a string and running the appropriate function.

References

[1] J. Li et al. “The fluxes of H_2O_2 and O_2 can be used to evaluate seed germination and vigor of Caragana korshinskii”. In: *Planta* 239.6 (2014), pp. 1363–1373.

[2] V. Ossipov et al. “Broad-specificity quinate (shikimate) dehydrogenase from Pinus Taeda needles”. In: *Plant Physiology and Biochemistry* 38.12 (2000), pp. 923–928.

[3] J. Van Asbrouck et al. “Predict emergence quality through oxygen consumption during the first hours of germination: Q2 and 'ASTEC values'”. In: 70^{th} *IIRB Congress* P3.12 (2007).

[4] L. W. Woodstock and D. F. Grabe. “Relationships between seed respiration during imbibition and subsequent seedling growth in zea mays L.” In: *Plant Physiology* 42.8 (1967), pp. 1071–1076.