# Predicting potato-plant vigor by parametric hyper-curve fitting

Atza, E.

**DOI**
10.4233/uuid:96a00229-3639-43a5-a91c-797cae0ba57f

**Publication date**
2025

**Document Version**
Final published version

**Citation (APA)**
Atza, E. (2025). *Predicting potato-plant vigor by parametric hyper-curve fitting*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:96a00229-3639-43a5-a91c-797cae0ba57f

# Predicting potato-plant vigor by parametric hyper-curve fitting

## *Proefschrift*

*ter verkrijging van de graad van doctor*
*aan de Technische Universiteit Delft,*
*op gezag van de Rector Magnificus Prof.dr.ir. T.H.J.J. van der Hagen,*
*voorzitter van het College voor Promoties,*
*in het openbaar te verdedigen op*
*dinsdag 3 juni 2025 om 12:30 uur*

*door*

Elisa ATZA

*Master of Science in Mathematics*
*Rheinische Friedrich Wilhelms-Universität Bonn, Duitsland*
*geboren te Oristano, Italïe*

*Dit proefschrift is goedgekeurd door de promotoren.*

*Samenstelling promotiecommissie:*

| | |
|---|---|
| Rector Magnificus | Voorzitter |
| Prof.dr.ir. C. Vuik | Technische Universiteit Delft, promotor |
| Dr. N.V. Budko | Technische Universiteit Delft, copromotor |

*Onafhankelijke leden:*

| | |
|---|---|
| Prof.dr.ir. G. Jongbloed | Technische Universiteit Delft |
| Prof.dr.ir. T. J. H. Vlugt | Technische Universiteit Delft |
| Dr. F. A. van Eeuwijk | Wageningen University and Research |
| Prof.dr. A. Micheletti | Universita' degli studi di Milano, Italie |
| Prof.dr. F. Vermolen | Universiteit Hasselt |

*An electronic version of this dissertation is available at:*

`http://repository.tudelft.nl/`

*Het onderzoek in dit proefschrift werd (mede) gefinancierd door HZPC B.V., Aveeris Seeds,* NWO.

To the many people who gave up comfort
and security to stand up for equality,
and to all my nieces.

# Contents

# Introduction

Agriculture and scientific research have been intertwined for hundreds of years, tracing back to the very notable example of Mendelian inheritance. For a long time the purpose of research in agriculture was to select for desirable traits and increase production. The last fifty years have seen a rising concern in society for sustainability and more environmentally-minded practices spurred by incontrovertible proof of climate change and a loss of biodiversity so staggering that it has been called a sixth extinction, [39], [9]. These concerns had a gradual but significant effect on consumer choices and have been reflected in policies regarding the environment and agriculture.

Currently, the European Green Deal and EU Commission mid-term Agricultural Outlook, [24], both stress the importance of innovative approaches to land use and crop management in order to reach European sustainability targets and preserve biodiversity. Indeed climate change is projected to interfere with current crop and land management systems whereas the amount of available arable land is not projected to increase. The challenges ahead indicate important research directions for researchers within and outside academia.

Many advances in agricultural research have been facilitated in the past decades by the popularization of high-throughput phenotyping technologies, i.e. devices and software which have made the processing of large amounts of images or biological samples faster and more affordable. This has culminated most recently in projects such as WatchITGrow (WIG) and the Netherlands Plant Eco-phenotyping Centre (NPEC) which allow practitioners, WIG, and academics, NPEC, to monitor plant development at an unprecedented level of precision.

The project "Flight to Vitality" that provided funding for this research was started in 2018 and involved a consortium of industry and academics: HZPC B.V., Averis seeds, TU Delft and Utrecht University. The goal was to combine high-throughput phenotyping technologies available in academia and industrial laboratories and commercial remote sensing services to connect the composition of the seed tuber and the vigor of the resulting potato plant.

Potato is the world's third most important crop for human consumption, as food and a source of starch [27]. Among carbohydrate-rich staple crops such as wheat, grain maize, oats, and rice, potato is particularly important because it has the highest yield per hectare thanks to water and space efficiency and therefore it is a crucial crop for facing the food demands in areas where availability of arable land is not expanding, [28].

The cultivation of potatoes presents unique challenges, most importantly because potato plants rely on vegetative propagation as planting potatoes at scale from true potato seeds presents difficulties. The production of certified seed tubers is closely regulated by national and international organizations, with the purpose of guaranteeing that the seed has been produced and multiplied in pathogen-free soil, is healthy and true to variety specifications, [48]. In the European Union, the Netherlands is the biggest producer of seed potatoes, with HZPC being a market leader [26], both thanks to favorable soil conditions and to a long history of innovation and improvement of the certified seed tubers. The quality and health of the seed tubers is one of the most important contributing factors to crop quality and yield. Awareness of this fact motivates ambitious collaborative projects, see, e.g., [38], which aim to increase the availability of high-quality seed potatoes in developing countries.

Even though the Netherlands is uniquely positioned in Europe for the production of seed potatoes due to its fertile and pathogen-free soil, producers have been registering a non-uniformity in the growth of seedlots of the same variety. This creates understandable friction between the seed-potato producers and their customers as non-uniformity of potato plants can lead to a diminished harvest. These difficulties are tied to the absence of an effective germination test for seed potatoes. Such a test, present for other crops, e.g., barley, is a cheap way to measure the vitality of a seedlot. Vitality, or vigor, is an important quality of plants, referring to several simultaneously occurring traits: good emergence, uniform growth, and the number of stems per plant. Strong vitality in the early stages of development usually implies a strong yield. Several studies have linked vigor of seed tubers to their storage conditions and physiological age, acknowledging also the importance of genotype [71], [40], [21].

The idea for the project "Flight to Vitality" (FtV) was sparked by the heuristic observation (HZPC, Averis Seeds B.V.) that the seed potatoes of the same variety (genotype, cultivar) originating from different production fields exhibit nonuniform growth pattern. This difference in the apparent vigor of seedlots of the same variety contradicts the goals of the seed-potato producers who aim to deliver the seeds of uniformly high quality, and may force the farmers to harvest plants some of which have not yet reached their potential yield.

The main goal of the FtV project was to identify the discriminating features of the seed tubers that cause the nonuniform emergence and growth rates of the potato plants. To achieve this ambitious goal the project consortium set up a multi-year experiment to analyze the seed tubers from different production fields and to monitor the plant growth, both in realistic test-field conditions and in controlled climate rooms. The latter were meant to mimic different climatic regions and potential stress factors affecting the growth of potato plants.

A variety of biochemical properties of the seed tubers could be measured in the laboratory thanks to high-throughput techniques such as Fourier Transform InfraRed (FTIR) spectroscopy, Hyper-Spectral Imaging (HSI), X-Ray Fluorescence (XRF), rRNA amplicon sequencing for the microbiome, and untargeted metabolomic mass spectroscopy. This wide spectrum of features was deemed to cover the main properties of the seed tubers that could potentially contribute to the observed difference in vigor.

To acquire a robust picture of the seed-tuber performance, field trials have been conducted during three consecutive years in three distinct geographic locations, two in the Netherlands and one the south of France. The growth of plants was monitored by unmanned-aerial-vehicle (UAV) imaging, with visual-spectrum (RGB) and multi-spectral cameras, during the period between 15 and 70 days after planting (DAP). Additionally, two climate rooms were built in Stiens (Netherlands), where potted plants have been grown in four climatic conditions, warm and wet, cool and wet, warm and dry, and cool and dry. The plant growth in the climate controlled rooms was monitored by a moving RGB camera setup starting at 15 DAP until 40 DAP, due to earlier closure of the the plant canopies compared to the field experiments.

The experimental design was updated in all environments after the first project year (2019) in order to make the measurement of plant vigor more accurate. In the years 2020 and 2021, a new drone operator was contracted for monitoring the field experiments, physical markers were placed in the field to be used for time-alignment, and a more stable rolling scaffolding was bought for the camera in the climate rooms. The author of this thesis was responsible for the extraction of the plant growth data from the acquired images and the development of a predictive model for the seedlot vigor in terms of the seed-tuber properties.

The quantification of the plant vigor from drone images is a challenging task, initially requiring the identification/definition of a quantity, measurable from an aerial photograph, which is a good indicator of such a complex trait. Due to the limited spatial resolution of the images, small size of plots, and a relatively large scale of the project, measurements of the plant biomass, number of stems per plant, and the plant height could not be achieved neither from the images nor directly on the ground. At the same time, we could reliably measure the average canopy coverage in each plot, thus being able to detect the nonuniformity of growth between the plots. Upon correcting for various image distortions and removing the growth trends due to the field inhomogeneity, the leaf-canopy-area data could be used as a dependent variable for regression.

In 2020 we explored associations present between the leaf-canopy data for 2019 and 2020 and the seed-tuber data available at that point (FTIR, XRF, HSI). To this end, we employed the chemometric standard method of regression – Partial Least-Squares (PLS). This method has been developed for highly collinear datasets, such as those usually arising in chemometrics, where the number of independent variables (predictors) is much larger than the number of studied samples (experiments). In our case, we studied 180 seedlots, and the spectra resulting from the FTIR spectroscopy alone contain 1400 features (spectral amplitudes). Therefore, we deal with an underdetermined linear system and in more general terms, an overparameterized problem. The preliminary results of this association study were presented at the ECMI 2021 conference and are published in the corresponding proceedings [3].

In a subsequent thorough analysis involving the the FTIR, XRF, HSI and microbiome data, we observed that the cross-validated complexity of the PLS regression models was very often equal to one, i.e., on average, only the first PLS component produced the smallest residual on the testing dataset. The PLS algorithm finds a solution to an underdetermined problem in a Krylov subspace of a given dimension. The choice of the dimension is part of the model calibration and regularization procedure, and for our purposes, this choice needed to be robust enough to be used on unseen datasets, hence, the decision to use cross validation. This indicated the limited applicability of the PLS method with the datasets at hand. In particular, a smooth solution (vector of regression coefficients) found in the Krylov subspace of dimension one did not offer much insight into the contribution of each individual feature to the model.

In March 2021 the microbiome sequencing for the years 2019 and 2020 was completed by the Plant-Microbe Interactions group at the University of Utrecht, which resulted in a joint paper [58]. In this work the regression was performed with the algorithm popular in the microbiome community – the Random Forest (RF) method. This method allows naturally encoding the taxonomy of the measured microbiota, but is highly nonlinear from the mathematical point of view, i.e., it implements a nonlinear association model between the tuber features (microbial population sizes) and the canopy area. The sophisticated nonlinear model behind the RF method, however, produced results very similar to the linear model implemented in the PLS method, which prompted us to question the applicability and choice between linear and nonlinear association models, especially, in the case of limited training datasets.

This initiated a theoretical study that eventually led to the publication [4] where we have revisited the topological structure of the multiple linear regression (MLR) model and showed that it is equivalent to fitting a hyper-curve parameterizable by a single scalar parameter. In particular, this approach allows viewing each individual predictor variable as a function of the dependent variable, hence, the Inverse Regression (IR) name of one of the possible realizations of our approach. This method is especially suited for severely overparameterized problems with diverse predictor data and has been further adapted by resorting to the Discontinous-Galerkin (DG) approach to accommodate for the possible discontinuities in the dependence of the predictor (tuber property) on the dependent variable (plant vigor) in the paper [5] and elaborated in Chapter 4. The IR method also allows analyzing the quality of the predictor data and systematically removing the features (predictors) that are either too noisy or do not comply with the linearity assumption of the MLR model. In [5] we have applied the DGIR method to the complete range of the seed-tuber data (with the exception of microbiome) and were able to narrow down the amount of seed-tuber information necessary to make a vigor prediction.

This thesis describes the FtV project and its results with the focus on the algorithms and mathematical techniques, and presents a detailed discussion of the PARCUR, IR and DGIR regression methods. We begin by outlining the research questions and previewing the main findings of the FtV project in Chapter 1.

In Chapter 2, we describe in detail the setup of the field and climate room experiments and the procedures we devised to measure the leaf-canopy area in both. For the measurement of plant vigor we had to rely on both existing and novel methods, and this chapter provides sufficient information about all techniques and algorithms for our results to be reproducible.

In Chapter 3, we introduce and explore the mathematical properties of the PARCUR and IR models. We give explicit conditions for the model to make exact predictions in an ideal case. We offer examples of how the method works on synthetic data and on a publicly available dataset of FTIR spectra for PET yarn.

In Chapter 4, we develop a DGIR version of the IR method and present the results of regression on the FTIR, XRF, HSI, and metabolomic datasets for all three project years. We develop both variety-agnostic and variety-specific regression models and identify seed-tuber features sufficient for a good prediction of the plant vigor.

In the Conclusions, we discuss the lessons learned from the FtV project with the focus on how mathematics can further contribute to the solution of some challenging problems in modern agriculture.

In Appendix A we include figures illustrating the model performance from 4 in all cases.

# Chapter 1

# Research questions and preview of results

Although this is a thesis on applied mathematics, since the main application area was horticulture, in the tradition of scientific articles on biology, the main results of the thesis will be summarized at the beginning. The research question of the project “Flight to Vitality” was formulated by the industrial partners, and one of the results of this thesis is the answer to this important practical question. The second research question has emerged while trying to solve the first one, and is of mathematical nature. A partial answer to that question is the second result of this thesis. Both the questions and the results are only briefly described in this chapter and are further elaborated in the subsequent chapters.

## 1.1 Is it possible to predict the vigor of potato plants from seed-tuber properties?

On the most basic level the research question of the “Flight to Vitality” project can be formulated as follows.

> Given the data about the chemical constitution and microbiome of the seed tuber, is it possible to predict the vigor of the potato plant?

Having carried out a three-year experimental campaign and having applied all the necessary data-processing and machine-learning (ML) techniques we arrived at the conclusion that the answer to this question is “sometimes, yes”. Specifically, we have been able to predict the vigor of one potato variety in all conditions fairly well, the

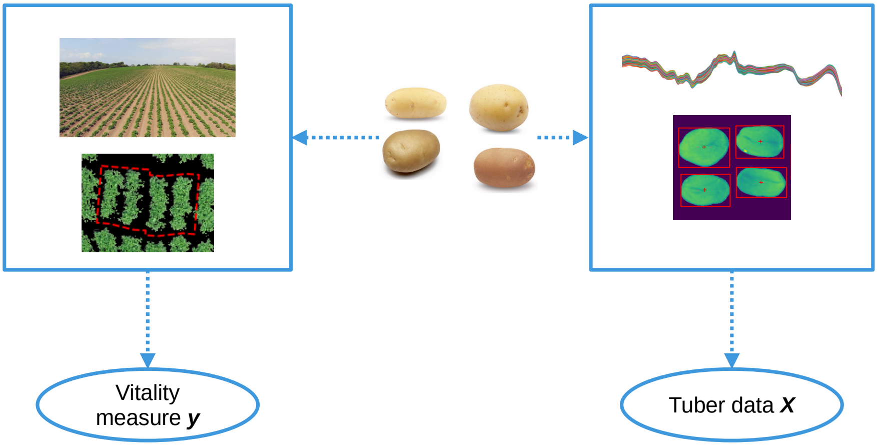


Figure 1.1: In the FtV experiment we study selected potato seed tubers in the lab and in the field. Several chemical and biological aspects are measured in the laboratory, these features constitute our matrix $X$ of tuber data. In the field, thanks to UAV imaging and some software specially developed for this project, we can measure the plant's vitality, which constitutes our dependent variable $\boldsymbol{y}$.

vigor of one other variety to a lesser extent (under certain environmental conditions), and failed to make a robust predictive model for the remaining four varieties.

Of course, there are many caveats to this answer. For example, why there had to be a variety-specific model and not one model that would be predicting the vigor of any potato variety? Such a variety-agnostic ML model was in fact created. However, it did not perform well at all. Subsequent analysis of the predictor data has shown that, for some predictors (spectral components, abundances of chemicals and metabolites, etc.), the mathematical form of association with the vigor parameter was very different between the varieties. This indicates that the vigor-enhancing properties or strategies may depend on the seed-tuber genotype.

Another warning is about the way the different environmental conditions were treated. It is well-known that different genotypes perform differently in the same environmental conditions. In our experiments, we have observed this G×E interaction when the early-emerging varieties were adversely affected by abnormally low temperatures in one field and left-over herbicide in another field. These effects were also observed on the level of seedlots. Thus, a proper predictive model should have included this genotype-specific reactions on the environment as well as the time-series of the environmental variables. While the latter time-series were available, we had no calibrated growth models for the genotypes in question at our disposal.

The lack of predictability of the vigor for certain varieties can thus have various explanations. On the one hand, there was a lack of consistency in the vigor of the seedlots between the fields for these varieties. Hence, the vigor of these varieties may simply be varying more randomly, independently of the seed-tuber origin. That immediately poses a follow-up question for biologists: what makes the Festien variety perform so consistently and predictably in various conditions? On the other hand, is there perhaps a better, more robust, way to measure the vigor, such that the 'unpredictable' varieties, that are often better performing on average, can still be predicted?

In the search for obvious and stable predictors of vigor, an over-abundance of data about the chemical composition and microbiome of the seed tubers was collected during this project. We have developed a special technique to characterize the predictive power of each individual predictor, be it a component of the infrared spectrum (spectral amplitude at a given wavelength) or a specific metabolite. Having applied this technique to all datasets with the exception of microbiome, we observed a surprisingly poor performance of the *a priori* most promising metabolome dataset. At the same time, the technique revealed that the relatively cheap spectroscopic data had the best predictive power. While this can be considered a good result from the practical/economic point of view, the reasons behind the lack of (linear) consistency in the metabolome data are not clear. This could be the sign of either a natural high variance in the abundances of metabolites in the seed tubers or a strongly nonlinear dependence of the vigor on the metabolite content.

## 1.2 How to solve an overparameterized regression problem?

While searching for a suitable machine-learning technique to answer the research question discussed above we came across the following mathematical question.

> Given $p$ possible predictors and $n$ training samples, where $n \ll p$, will a multiple linear regression model be able to make good predictions?

This question is especially important when a multiple linear regression (MLR) model fails to make a good prediction of the testing dataset. Is it because the underlying biological association is weak, or the training dataset is somehow incomplete, or is the assumption about the (multiple) linear relation between the predictors and the dependent variable too restrictive? If the latter is true, then one may have to consider a more sophisticated nonlinear ML model, which, however, may also turn out to be far less interpretable. Moreover, having additional, e.g., nonlinear, degrees of freedom in an ML model does not come for free, as 'learning' the coefficients of such a model may require more training data. Does it make sense to switch to a highly overparameterized non-linear model when the linear model is already overparameterized, i.e., $n \ll p$, and there is no possibility to increase the size $n$ of the training dataset?

We have obtained partial answers to these questions. Yes, a well-trained MLR model will be able to make good predictions even in the presence of nonlinear dependencies of the dependent variable on the majority of the predictors. In fact, it is enough to have a few 'proper' predictors in the dataset to achieve this. In the ideal (noiseless) case, there is even no need to know which of the predictors are good and which are bad. While this sounds very convenient, such a model would offer an illusion of understanding of the underlying association, [46]. Hence, here and especially in the realistic case with noisy data, it is prudent to remove the 'improper' predictors, which incidentally also significantly improves the model performance.

To arrive at these answers, we had to reformulate the MLR model in a way that made its topology more clear. The term 'linear', which stems from the algebraic structure of the model equations, makes one think that the relation between the predictors and the dependent variable is a multi-dimensional linear object in the $(p+1)$-dimensional space. Indeed, an equation of the plane, e.g., in three dimensions, has the form:

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 = 0, \tag{1.1}$$

where $(x_1, x_2, x_3)$ are the three coordinates, and $\langle \beta_1, \beta_2, \beta_3 \rangle = \boldsymbol{n}$ are the components of the vector $\boldsymbol{n}$ normal to the plane.

The algebraic relation behind the multiple linear regression has the form:

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p = y, \tag{1.2}$$

where $x_j$ are the predictors, $y$ is the dependent variable, and $\beta_j$ are the unknown coefficients that must be learned from the training data. The equation of the plane (1.1) can be rewritten in a similar form:

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 = y, \tag{1.3}$$

where $y = -\beta_3 x_3$ plays the role of the dependent variable. Hence, it appears that the data points $(x_1, \dots, x_p, y)$ of the multiple linear regression model are supposed to lie on a hyper-plane in a $(p+1)$-dimensional space.

While this is undoubtedly true, we were able to show that the points of the trained overparameterized MLR model also lie on a hyper-curve that can be parameterized by a single scalar parameter $s$. A hyper-curve is completely defined by $p+1$ scalar functions of this parameter:

$$y(s), \;\; x_j(s), \;\; j = 1, \dots, p. \tag{1.4}$$

Moreover, in many cases the parameter of the curve may be chosen as $s = y$, i.e., the dependent variable. The latter choice allows to consider each predictor as a function of the dependent variable and eventually delivers the tool for discarding improper predictors from the model.

By shifting the focus onto the behavior of predictors as functions of the dependent variable, the parametric hyper-curve model offers the possibility to choose the model of the functional relation sought between these variables. This relation may be continuous, e.g., polynomial, but it may also be piecewise-continuous or even discontinuous (a broken hyper-curve). To accommodate for the datasets that correspond to the latter two cases, we develop a finite-element projection procedure for deriving the discrete ML model, opening an interesting avenue of research at the interface of numerical methods and machine learning. Finally, this approach offers an alternative way of tackling the difficult problem of noisy predictors, [36], by regularizing the functional representation for each predictor individually.

## 1.3 What can be improved?

### 1.3.1 Climate-controlled room experiments

In principle, acquiring experimental data in a climate-controlled room is an excellent idea. The field experiments in 2020 and 2021 have shown that even minor changes in climatic conditions at key stages may and do strongly affect the outcome of the experiment. Therefore, validating the results of field experiments and/or extending them to new conditions in climate-controlled rooms would be very helpful. However, growing potatoes in containers, data acquisition, and data processing presented certain challenges that prevented us from realizing the full potential of climate-controlled rooms in the FtV project.

First, the data acquisition was disproportionately costly in terms of the manual labor and time spent by the operator taking the pictures. This process can and should be automated by employing a fixed multiple-camera setup, rather than a moving single camera. The latter option was realized in the FtV project and required the development of a trivial but convoluted software for extracting the leaf-canopy data that depends on the particularities of the moving-camera setup and is difficult to generalize and deploy in practice.

Second, the standard techniques for removing the spatial trends from the data are not applicable to individual containers that were used to grow the potato plants in climate-controlled rooms. In this case, the trend, e.g., due to a non-uniform irrigation, is a discontinuous function of space coordinates and a new approach, specifically suited for container-based experiments, must be developed. Also, a smaller number of replicates, necessitated by the spatial constraints of the rooms, must be taken into account.

Third, the small size of containers significantly affects the growth of the plants, convoluting the correspondence between the field tests and the experiments in climate-controlled rooms. Further biological, mechanistic modeling is required to elaborate this correspondence and incorporate it into the data-processing pipeline.

### 1.3.2 Multi-spectral drone data

In all three years, during the field tests, apart from the RGB camera, the drone was also carrying a multi-spectral camera that took images of the same field in three near-infrared frequency bands. This information can be used, e.g., to estimate the chlorophyll content of the leaves, which could also be related to the plant vigor. While recently, commercial drone operators have started implementing the necessary measurements (e.g., GPS coordinates) and pre-processing steps in their software, in the FtV project, the multi-spectral images could not be time-aligned (transformed to the same frame of reference on each measurement date) with the RGB orthophotographs. Therefore, the plot-identifying polygons found in the RGB images could not be reused with the multi-spectral images. Whereas, recreating the plot polygons and the time-alignment directly on the multi-spectral images failed due to significantly lower image resolution and the apparent invisibility of the ground-based markers in the near-infrared spectrum.

### 1.3.3 Plant-level data

We developed an effective algorithm that detects plot boundaries for seedlot plots of approximate size 3 by 4 meters, as shown in Figure 1.2, and can measure canopy ridge-wise with a semi-automated canopy segmentation procedure. Although this procedure does require some manual supervision and intervention, the segmented plot boundaries could then be easily transferred onto the orthophotographs of the test fields taken at different dates. We only had to detect the polygonal plot boundaries in a suitable image taken on some selected date and use the field markers visible in all images to transform the coordinates of the vertices to the reference frame of other orthophotographs.

The resolution of orthophotographs in the FtV project was fine enough to estimate the canopy area on the sub-plot (ridge) level. Knowing the number of plants in the ridge, we could estimate the average canopy per ridge. This gave us the potentially valuable information about the variance of the seedlot canopy area. However, as there were only four ridges inside each plot, the actual plot variance was significantly underestimated by the variance of the ridge data. Hence, the next step in the processing of the RGB drone data is to extract the canopy area of individual plants, e.g., with a parallel-moving ridge-conformal window searching for gaps between the plants. Apart from the possible need to increase the image resolution, there are many other challenges on this path. For example, in the images taken at the earlier dates, it is hard to distinguish the emerging plants from the stones and cracks in the soil. Whereas, in the images taken at the later dates, it is difficult to distinguish between the joint canopies of two separate plants and the important case where a plant has failed to emerge. A way forward is to increase the temporal resolution by taking the images more often, and to incorporate the dynamic information into the plant segmentation procedure.

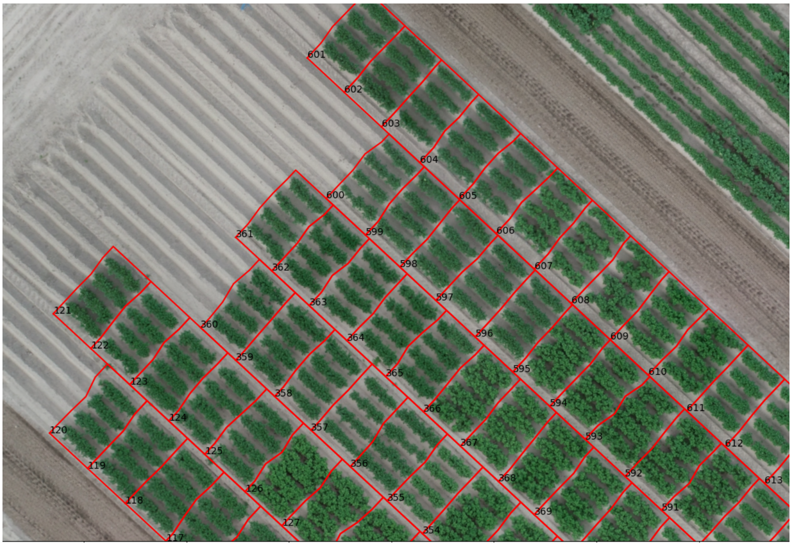


Figure 1.2: The figure shows the polygonal boundaries detected in the field of Kolummerwaard, in black is the plot number from the field scheme for seedlot association.

## 1.3.4 Modeling environment and new genotypes

The main hypothesis of the FtV project was the possibility to predict the vigor of potato plants from the properties of the seed tubers. This is a data-based approach and we have shown that such a prediction is possible to some extent. A variety that is less vigorous, but more stable across different environmental conditions was much better predictable than the other varieties. The "unpredictable" varieties reacted much more strongly on the adverse conditions, such as residual herbicide and early cold temperatures, but are also generally more vigorous, i.e., grow more rapidly and have larger canopies.

We believe that the predictive ability for the vigorous varieties could be improved, if the models were augmented with the environmental data, e.g., temperature, air and soil humidity, Nitrogen content, etc. As a source of inspiration for this new type of data-based model, we suggest to look into the mechanistic plant growth models, [51], such as WOFOST [62], [13], which codify the response of a variety to various environmental factors. However, it also clear that the mathematical structure of these mechanistic models will have to be revisited first, with the goal to reduce the number of tuning parameters.

In the FtV project we have considered six different potato varieties and come to the conclusion that the vigor-predictive model should be variety specific. This raises the natural question on how to extend these models to other genotypes. Does it mean that the costly experiments similar to the ones in the FtV project will have to be performed with each new variety? While we believe that additional experiments are, probably, inevitable, one should seek a data-based explanation of the model coefficients in terms of the genetic information, such as Single Nucleotide Polymorphisms that distinguish one variety from another. Hopefully, with enough experimental data and a suitable machine-learning approach, a mathematical pattern will be uncovered that allows to predict the tuber-vigor connection even for previously unseen or yet to be bred varieties.

# Chapter 2

# Experiments and data extraction

While academic mathematicians were not involved in the initial formulation of this project and the design of its experiments, mathematical methods played substantial role in the subsequent stages, namely, in the extraction of plant-growth data, the definition of the vitality measure, and the creation of a predictive machine-learning model.

The measurement of plant features, e.g., the size, shape and physiological state of a leaf, is called *phenotyping* and represents both a well-studied subject and a rapidly developing research direction, [18], [29], [44]. It involves the methods from computer vision and photogrammetry, [42], [2], and has recently been reinforced by the techniques of Artificial Intelligence (AI), [57].

In this project, the main plant feature of interest was the size of the leaf canopy. Since the source of the data about the canopy were the RGB images of the test field acquired by a drone-mounted camera, several intermediate data-processing steps have been performed to extract the desired canopy data.

This chapter describes the design of experiments, and elaborates both the standard methods and the techniques specifically developed in this project, pertaining to the extraction and spatial correction of the vitality data. All these techniques have been published in the Protocols section of [7] and in the expanded and updated Protocol section of [6].

## 2.1 Project hypothesis and design of experiments

Seed potato production aims to breed varieties or genotypes that exhibit one or multiple desirable traits. Potatoes are used in many ways by the food industry, as an ingredient for direct consumption, and as a source of starch. This division is reflected in the breeding, for example Averis Seeds focuses on the production of starch-rich potato varieties.

The focus of this research are seedlots or batches. A seedlot is a group of seed tubers of the same variety with the common origin, i.e., the production field, the producing farmer, and the storage conditions. In the subsequent growing season, potato plants grown from the seedlots of the same variety exhibit variations that can not be explained by genetic properties, since the seedlots of a variety contain genetically identical tubers (clones). These variations have been observed for many years with different potato varieties and in different growth conditions. Therefore, the project's main hypothesis is that the ability of a seed tuber to produce a healthy plant (vitality of a tuber) depends on the origin of the seed tuber and that this variation in vitality may happen across a variety-specific range. Another assertion of the project is that the origin of the tuber is imprinted in its bio-chemical properties. Thus, one should be able to predict the vitality of the potato plant by analyzing the chemistry and the microbiome of the seed tuber.

To validate this hypothesis, a three-year field experiment was carried out with a set of six varieties of varying average vitalities. Each variety was represented by 30 seedlots, amounting to 180 seedlots per year. The seed potatoes involved in the experiment were cultivated in the production fields all over the Netherlands with varying soil and crop management histories and were subject to possibly different storage conditions.

To confirm the consistency of the vitality variations between the seedlots, the same seedlots were simultaneously grown in different natural and artificially controlled environments. The natural variations were achieved in the years 2019, 2020, and 2021 by planting in the following three test fields:

1. Montfrin (M), in the South of France (54.4980 N, 5.1090 E)
2. Kollumerwaard-SPNA (S), in the North of the Netherlands (70.4325 N, 6.9825 E)
3. Veenklooster (V), in the North of the Netherlands (70.3935 N, 6.7080 E)

The artificial environments were created in two climate-controlled rooms located at the HZPC facilities in Stiens. The two rooms were kept at different temperatures, one at 10°C, and another at 20°C, and each room was subdivided into a wet and a dry zone according to the levels of applied irrigation. Thus the climate rooms implemented four different conditions: warm-wet, warm-dry, cold-wet, and cold-dry. Three experiments were conducted in the rooms between April 2019 and August 2020.

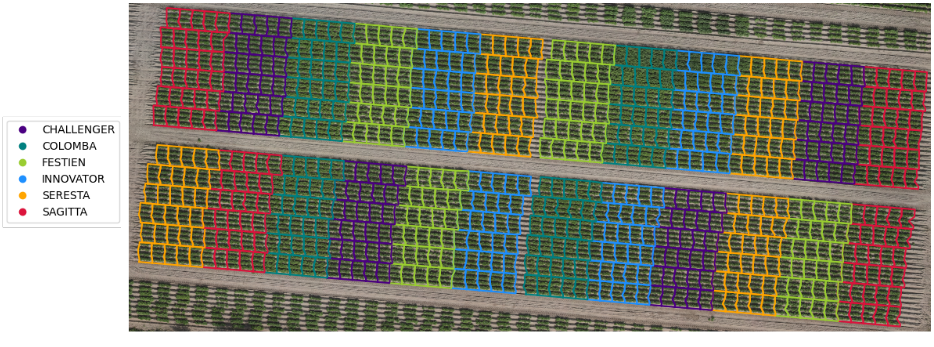


Figure 2.1: Randomized complete block design of the Veenklooster test field in 2021. Each genotype is repeated four times as four randomly located compact blocks (colors). Each seedlot has one plot (small polygons) randomely located inside the corresponding genotype block, i.e., there are four plots of each seedlot in total.

As mentioned above, a single field trial involved 180 seedlots belonging to 6 varieties (30 seedlots per variety). The seedlots were planted according to a randomized complete block design (RCBD). Each seedlot was represented by 96 seed tubers. These tubers were divided into four replicating *plots*, each containing 24 tubers. Thus, with 180 seedlots, this resulted in 720 plots distributed over the test field. As potatoes are planted inside the elevated soil structures called ridges, each seedlot plot stretched across four neighboring ridges containing 6 tubers of this seedlot planted along (inside) each ridge. This planting arrangement was achieved by manually placing the tubers into the ground from a tractor while it was creating the ridges, which sometimes resulted in a significant irregularity of the plot boundaries as can be seen in the experimental design of the Veenklooster test field in 2021 shown in Figure 2.1.

In the climate room trials $11 \times 11 \times 12$ cm containers were filled with sand and hosted a single seed tuber each. These containers were placed on large tables in a grid-like pattern along 10 columns and 72 rows, implementing a randomized complete block design. Studying the emergence of potato plants in such small containers was an untested approach. Unfortunately, many of the potato plants have failed to emerge in these experiments and the canopies of individual plants that did emerge tended to cross the boundaries of the corresponding containers, complicating the task of measuring the canopy area. Therefore, the data from the climate room trials were eventually discarded. Yet, some of the phenotyping techniques that were developed for the climate-controlled room setup may be re-used in future modified experiments.

The development of potato plants over time was monitored by taking the RGB and multi-spectral images of the complete field and climate-room tables at certain moments after the planting of the seed tubers up until (and sometimes also after) the canopy closure, i.e., the moment when the leaf canopies of the neighboring plants begin to overlap. The images of the test fields were acquired with a drone-mounted camera, whereas the images of the climate-room tables were acquired with a camera mounted on a frame manually positioned by a human operator. The dates of the drone images of the test fields can be found in Table 2.1.

Although the same 6 genotypes were studied in all years, the seedlots that represented these genotypes had a different origin between the years. The idea was to train a machine-learning model explaining the variations in vitality between the seedlots on the experimental data from one year and test it on the data from another year.

For each seedlot, a vast range of seed tuber features has been measured at the HZPC facilities in Metslawier and at the microbiology department of the Utrecht University. Specifically, the following seed-tuber data was acquired:

1. Infrared spectra of frieze-dried samples with the Fourier-Transform Infrared-Spectroscopy technique (FTIR)

2. Near-infrared spectral images of the wet tuber slices with Hyperspectral Imaging camera (HSI)

3. Abundances of 11 elements in the frieze-dried samples with the X-Ray Fluorescence technique (XRF)

4. Abundances of small molecules in frieze-dried samples with the Quantitative Time of Flight Mass-Spectroscopy technique (Metabolome)

5. Abundances of microbial and bacterial species with the micro-array technique based on sequencing and identifying of the genetic material on the peeled skin and in the 'eyes' of the tubers (Microbiome)

The seed lots were collected from the various seed growers and stored at 8 °C. From each seed lot 4 batches of 5 tubers were washed, dried and peeled. Both the flesh and peel fraction was cut into small pieces using ceramic knives and frozen in liquid nitrogen. Two 50 ml tubes were filled with frozen potato flesh and peel respectively, weighed and stored at -80 °C before freeze drying. A separate 150 ml container was filled with  100 g frozen potato flesh, weighed and stored at -20 °C before freeze drying. All samples were weighed after freeze drying and dry matter percentages were calculated.

### XRF data

From the 150 ml container, 8.3 grams of freeze dried potato flesh was milled in a zirconium oxide grinder and sieved. A mixture of 8.0 g powder and 2.0 g binder was well homogenized and pressed into a pellet. The dry weight concentrations of Mg, K, Ca, Fe, Cu, Zn, P, S, Cl and Mn were determined using energy dispersive X-ray fluorescence spectroscopy (Bruker S2 PUMA) and converted to concentrations in fresh weight using the dry matter percentage.

| Field | Date | DAP | 0-plots (%) |
|---|---|---|---|
| M | 19-04-10 | 36 | 626 (86.9%) |
| M | 19-04-19 | 45 | 20 (2.78 %) |
| M | 19-04-26 | 52 | 0 (0.0 %) |
| V | 19-05-24 | 36 | 44 (6.1 %) |
| V | 19-05-29 | 41 | 0 (0.0 %) |
| V | 19-06-07 | 50 | 0 (0.0 %) |
| S | 19-06-07 | 36 | 4 (0.56 %) |
| S | 19-06-19 | 48 | 0 (0.0 %) |

| Field | Date | DAP | 0-plots (%) |
|---|---|---|---|
| M | 20-04-10 | 35 | discarded |
| M | 20-04-13 | 38 | 199 (27.6%) |
| M | 20-04-16 | 41 | 38 (5.3 %) |
| M | 20-04-18 | 43 | 13 (1.8 %) |
| M | 20-04-22 | 47 | 0 (0.0 %) |
| M | 20-04-25 | 50 | 0 (0.0 %) |
| V | 20-05-27 | 35 | 1 (0.14 %) |
| V | 20-05-30 | 38 | 0 (0.0 %) |
| V | 20-06-07 | 46 | 0 (0.0 %) |
| V | 20-06-10 | 49 | 0 (0.0 %) |
| V | 20-06-12 | 51 | 0 (0.0 %) |
| S | 20-06-03 | 35 | 7 (0.97 %) |
| S | 20-06-10 | 42 | 0 (0.0 %) |
| S | 20-06-12 | 44 | 0 (0.0 %) |
| S | 20-06-15 | 47 | 0 (0.0 %) |
| S | 20-06-19 | 51 | 0 (0.0 %) |

| Field | Date | DAP | 0-plots (%) |
|---|---|---|---|
| M | 21-04-02 | 24 | 26 (3.61 %) |
| M | 21-04-07 | 29 | 7 (0.97 %) |
| M | 21-04-08 | 30 | 3 (0.42 %) |
| M | 21-04-16 | 38 | 29 (4.03 %) |
| M | 21-04-19 | 41 | 3 ( 0.42 %) |
| M | 21-04-22 | 44 | 1 (0.14 %) |
| M | 21-04-26 | 48 | 0 (0.00 %) |
| M | 21-05-02 | 54 | 0 (0.00%) |
| M | 21-05-04 | 56 | 0 (0.00%) |
| M | 21-05-08 | 60 | 0 (0.00%) |
| M | 21-05-11 | 63 | 0 (0.00%) |
| M | 21-05-14 | 66 | 0 (0.00%) |
| M | 21-05-19 | 71 | 0 (0.00%) |
| V | 21-05-20 | 29 | 13 (1.81%) |
| V | 21-05-24 | 33 | 5 (0.69%) |
| V | 21-05-28 | 37 | 0 (0.0 %) |
| V | 21-05-31 | 40 | 0 (0.0%) |
| V | 21-06-04 | 44 | 0 (0.0%) |
| V | 21-06-07 | 47 | 0 (0.0%) |
| V | 21-06-11 | 51 | 0 (0.0%) |
| V | 21-06-18 | 58 | 0 (0.0%) |
| V | 21-06-23 | 63 | 0 (0.0%) |
| S | 21-06-18 | 15 | 653 (90.69%) |
| S | 21-06-23 | 20 | 31 (4.31 %) |
| S | 21-06-28 | 25 | 0 (0.0 %) |
| S | 21-07-03 | 30 | 0 (0.00%) |
| S | 21-07-05 | 32 | 0 (0.00%) |
| S | 21-07-08 | 35 | 0 (0.00%) |
| S | 21-07-14 | 41 | 0 (0.00%) |
| S | 21-07-16 | 43 | 0 (0.00%) |

Table 2.1: The dates (year-month-day) of the drone images of the three test fields (M, V, and S) in 2019 (left-top), in 2020 (left-bottom), and in 2021 (right). The 'DAP' column shows the time of the drone image in Days After Planting (DAP). The column 'zero plots (%)' gives the number of plots with no measurable canopy and their fraction among all plots in the field. The measurement on 35 DAP at Montfrin in 2020 was discarded due to unreliable segmentation. Measurements highlighted in green are used for regression.

### FTIR data

The freeze dried potato flesh of the 50 ml tube was grinded to a fine powder and homogenized. The powder of each batch was put into 3 wells of a 96-well sample plate and measured in a high-throughput FTIR spectrometer (Bruker Tensor II + HTS-XT). Spectral outliers among the replicates were identified and the corresponding samples were remeasured.

### HSI data

From each seedlot, approximately 50 tubers were selected and a 1 cm wide longitudinal slice from the center of each tuber was cut. The slices were placed on a moving platform and scanned using a push broom SWIR hyperspectral camera (900 – 2500 nm, Specim). Before each scan, a dark and white reference was scanned and used to convert the data to absorbance units.

The scans have been further processed as described in Section 2.9 in order to obtain the spectral measurements for pith and cortex.

### Metabolome

From both sample types (flesh and peel) untargeted metabolic profiles were measured in an incomplete block design following guidelines from the Metabolomics Quality Assurance and Quality Control Consortium (mQACC) [17]. Pooled samples were created by mixing 60 random lots, 10 per variety, for flesh and peel separately. These samples were used in separate dilution series and as precision references every 10 sample injections in the experiment runs.

The freeze dried sample (100 mg) was weighed in a 2 ml Eppendorf tube, 1.3 ml of methanol was added, vortexed and shaken for 30 minutes. After centrifugation (14000 rpm, 5 min.), 1 ml supernatant was transferred to a new tube and dried using a vacuum concentrator (1 mbar, 35 °C,150 min.). The residue was redissolved in cyclohexane (200 µl), milliQ water (300 µl) was added, shaken (10 min.) and centrifuged (14000 rpm, 10 min.). Using extended length tips, 180 µl of the lower aqueous phase was transferred to a 0.2 µl PVDF filter plate and centrifuged (1200 rpm, 4 min.).

The filtered samples were analyzed using a Waters Acquity I-class UPLC coupled with a Waters Xevo G2-XS QTOF MS. Chromatographic separation was achieved on a reverse-phase Acquity UPLC HSS T1.8 µm (2.1 x 100 mm) column (Waters) at 40°C. The mobile phases employed were A: water (MilliQ) containing 0.1% formic acid (UPLC grade, BioSolve) and B: acetonitrile (UPLC grade, BioSolve). The flow rate was maintained at 0.5 mL/min, and the injection volume was 5 µL. A gradient elution was employed, starting with 99% A for the first minute, followed by a linear gradient to 30% A over the next 10 minutes. A cleanup step was performed at 99% B for 1 minute, followed by 3 minutes of re-equilibration at the initial conditions. The total runtime was 15 minutes. The ESI source was operated in both positive and negative ionization modes with the following settings: capillary voltage, 3.00 kV; cone voltage, 40 V; source temperature, 120°C; desolvation temperature, 350°C; gas flow rate, 800 L/h (N2); cone gas flow rate, 50 L/h (N2). Leucine enkephalin was used as a Lock Spray reference.

Mass data were collected over the m/z range of 50-1200. MSe (Multiplexed Selected Ion Monitoring) was employed to acquire both parent ions and fragmentation data in a single run. The collision energy ramp was set from 10 to 40 V, allowing for the fragmentation of precursor ions across a range of energies.

Alignment of the chromatograms and peak picking was performed in Progenesis QI. The precision reference results were used to perform batch and drift correction for each individual peak in the nPYc-Toolbox [55]. Feature selection was done based on the linearity in the dilution series and relative standard deviation in the precision references.

## 2.2 Standard techniques

In this section we describe the standard tools used in the project: software for the orthophoto generation from drone images, and software for the removal of spatial effects from canopy measurements.

### 2.2.1 Orthophoto generation

During flight, drone-mounted cameras take hundreds of pictures along a pre-programmed path over the experimental field. These pictures are taken at a specific height of flight (30 meters in FtV) and are redundant, meaning that multiple pictures portray the same region in the field from different perspectives, in other words, they are overlapping. These images are not meant to be used as is. However, the stable height of the flight path and the redundancy of images are the features that facilitate the creation of the so-called orthophotograph of the entire scene. An orthophoto is a single picture of the full experimental field that portrays an orthogonal projection of each plant on the horizontal plane, i.e., as if seen from above at an infinite distance. An orthophoto makes the geometric measurements made from the image at any two different locations in the field comparable.

The creation of an orthophoto involves techniques of photogrammetry: feature extraction, 3D point-clouds, digital elevation models, etc. The quality and resolution of the original overlapping images and the flight-path stability influence the quality of the resulting orthophoto. For example, too little overlap between images leads to an excessive use of interpolation during the creation of the 3D point-cloud, which results in the presence of artifacts in the orthophoto. The creation of an orthophoto from drone images is required in many applications and several open-source and commercial solutions were available in 2019, at the start of our project.

In 2019 we obtained orthophotos from UAV imaging with the open-source WebODM software (Version 1.1.3) with the input parameters specified in Table 2.2.

| Parameter | Value | Effect |
|---|---|---|
| custom setting | high-resolution | |
| mesh-octree-depth | 12 | |
| min-num-features | 20.000 | |
| texturing-nadir-weight | 0 | |
| orthophoto-resolution | 1.0 | cm - pixel for the orthophoto |
| dem-resolution | 1.0 | cm - pixel for the elevation model |
| ignore-gsd | true | |
| build-overviews | true | |
| crop | 0 | orthophoto is not cropped |
| camera-lens | brown | |
| skip-3dmodel | true | |
| depthmap-resolution | 2000 | |

Table 2.2: WebODM input parameters

| Parameter | Value |
|---|---|
| Software name | Agisoft Metashape Professional |
| Software version | 1.6.0 build 9925 |
| OS | Windows 64 bit |
| RAM | 127.69 GB |
| CPU | Intel(R) Core(TM) i9-9900X CPU @ 3.50GHz |
| GPU(s) | GeForce RTX 2080 Ti |

Table 2.3: Software used to produce stitched orthophotographs of the fields in 2020

In 2020 the orthophotos were provided by the drone operator (Aurea Imaging), according to industry standards, and were obtained with the commercial Agisoft package, see Table 2.3, with the input parameters set as in Table 2.4.

### 2.2.2 Spatial effect removal

The average plot canopies obtained after the time alignment of images, removal of artifacts, plot localization, and segmentation, described below in Section 2.3, constitute the raw data and cannot be used to estimate of the mean seedlot canopy (by taking the average of plot repetitions), since the test fields are usually spatially non-uniform, which may systematically increase or decrease the measured canopy size in certain parts of the field.

Field inhomogeneities are a well known source of noise and bias in agricultural research and methods to minimize this error both by experimental design and by modeling have been studied since the 1930's. Indeed, the presence of the so-called *field effect* is one of the reasons for which agricultural trials are planted according to specific randomized designs. In our case, the trial fields were planted according to the randomized

| **Parameter** | **Value** |
|---|---|
| Coordinate system | W GS 84 (EPSG::4326) |
| Rotation angles | Yaw, Pitch, Roll |
| Point Cloud | |
| Points | 218,658 of 237,470 |
| RMS reprojection error | 0.202082 (0.684417 pix) |
| Max reprojection error | 0.607143 (18.1591 pix) |
| Mean key point size | 2.90074 pix |
| Point colors | 3 bands, uint8 |
| Key points | No |
| Average tie point multiplicity | 8.24931 |
| Alignment parameters | |
| Accuracy | High |
| Generic preselection | No |
| Reference preselection | Source |
| Key point limit | 40,000 |
| Tie point limit | 4,000 |
| Guided image matching | No |
| Adaptive camera model fitting | Yes |
| Depth maps generation parameters | |
| Quality | High |
| Filtering mode | Moderate |
| Dense Point Cloud | |
| Points | 185,097,287 |
| Point colors | 3 bands, uint8 |
| Depth maps generation parameters | |
| Quality | High |
| Filtering mode | Moderate |
| Model | |
| Faces | 84,330 |
| Vertices | 42,722 |
| Vertex colors | 3 bands, uint8 |
| Surface type | Height field |
| Source data | Sparse cloud |
| General | |
| Interpolation | Enabled |
| Strict volumetric masks | No |
| DEM | |
| Coordinate system | W GS 84 / UTM zone 31N (EPSG::32631) |
| Source data | Dense cloud |
| Interpolation | Enabled |
| Orthomosaic | |
| Coordinate system | W GS 84 / UTM zone 31N (EPSG::32631) |
| Colors | 3 bands, uint8 |
| Blending mode | Mosaic |
| Surface | DEM |
| Enable hole filling | Yes |
| Software version | 1.6.0.9925 |

Table 2.4: Agisoft input parameters

complete-block design, which randomly distributes the seedlots over compact variety blocks and distributes the repetitions of such variety blocks randomly across the field. Planting schemes can be very effective in mitigating the field-induced distortion of the measurements but, as shown in [14], particularly in large-scale experiments, the adoption of a good experimental design does not eliminate the need for additional accurate spatial modeling.

Indeed, even though our fields have been carefully prepared for the experiments, the spatial effect is visible to the naked eye, as is well-illustrated in Figure 2.4, showing the four plots of the same batch, with the plot depicted in the right-most image having significantly smaller canopies than the other three plots. Obviously this plot has been affected by some unfavorable growth conditions and would pull down the estimate of the mean canopy if included in the calculations as it is.

```r
library(statgenSTA)
library(SpATS)

CanopyMeasurement <- read.csv("/Path_to_Canopy_measurement.csv",
                               header = TRUE, sep = ",")

TableToFit <- createTD(data = CanopyMeasurement, genotype = "Batch",
                       repId = "Block", subBlock = "Block",
                       rowCoord = "Row", colCoord = "Col")

FittedModel <- fitTD(TD = TableToFit, trials=’trialname’,
                     traits = "Canopy", design = "rcbd")

BLUEsTable <- extractSTA(STA = FittedModel, what = "BLUEs",
                         keep = "Variety")
```

Listing 2.1: Code to produce the array of spatially corrected canopies from the raw measured array in R

Historically, correction for spatial inhomogeneities was, probably, first introduced by Papadakis in 1937 [50], and Bartlett in 1938 [10] as the so-called Nearest-Neighbor Adjustment (NNA), which connects the measurement $y_i$ at the location $i$ and the measurement at its neighbor or neighbors. The basic assumption of most methods implementing the NNA is that some degree of data-differencing removes the locally-linear growth trend [12]. In the various realizations of the NNA approach, one has to choose between the first and the second differences, but the quality of the corrected data depends on the assumption that the resulting residuals are stochastically independent, i.e. one needs to know or assume the level of spatial trend affecting the measurement. The NNA approach was revisited in the late seventies when the advances in computer technology made the data processing more accessible, [8].

Mathematically, the correction/representation for the measurement at the location $i$ via the measurement at $i - 1$ is of the form:

$$y_i = \rho y_{i-1} + e_i,$$

where the factor $\rho$ is assumed to be very close or equal 1, and the $e_i$ is a residual uncorrelated with $y_{i-1}$. This model is an example of an auto-regressive model of order one, and the correction is based on the retrieval of spatial variance-covariance structures from the data. In this way one can recover locally linear trends, but it is not possible to include any variety information, and missing data points cannot be easily handled.

The NNA model was expanded, [12], to include treatment information for an experiment with $n$ plots and $p$ treatments, with the plots planted in a long one-dimensional column:

$$\boldsymbol{y} = \gamma\mathbf{1} + X\boldsymbol{\tau} + \boldsymbol{e},$$

where $\boldsymbol{y} \in \mathbb{R}^n$ is the vector of plot measurements, $\gamma \in \mathbb{R}$ is a global effect, $\mathbf{1} \in \mathbb{R}^n$ is a vector of ones, $X \in \mathbb{R}^{n\times(p-1)}$ is the design matrix of the effects, $\boldsymbol{\tau} \in \mathbb{R}^{(p-1)}$ is the vector of fixed effects, and $\boldsymbol{e} \in \mathbb{R}^n$ is the vector of random spatial effects. A connection between this linear mixed-effect model and the ANN approach is derived from the formulation of the ANN method as the model $D\boldsymbol{y} = \boldsymbol{\epsilon}$, where $D$ is a differencing matrix, and $\boldsymbol{\epsilon}$ is the uncorrelated random noise. However, the need and amount of data-differencing has become a point of dispute in subsequent work [69] (as cited in [31]).

Another approach to modeling spatial variation without explicit differencing was introduced in [72] and [31] for $n$ plots organized in a $r \times c$ grid. Importantly, this formulation aims to take into account two sources of spatial variation: local and global, which are modeled not by estimating the variance-covariance structures, but by fitting a smooth function of the spatial variables. The residuals are assumed to be separable and second-order stationary so that their covariance can be represented as a function of two variables. The user has to choose the covariance functions empirically, with the aid of cross validation. All model parameters, including the two auto-regressive (AR) coefficients, one for the rows of the grid and one for the columns, are estimated by (restricted) maximum likelihood (REML). This approach admits the following mixed-model formulation:

$$\boldsymbol{y} = X\boldsymbol{\tau} + Z\boldsymbol{u} + \boldsymbol{\xi} + \boldsymbol{\eta},$$

where $\boldsymbol{y} \in \mathbb{R}^n$ is the vector of plot measurements, $X \in \mathbb{R}^{n\times t}$ is the design matrix of the fixed effects, $\boldsymbol{\tau} \in \mathbb{R}^t$ is the vector of fixed effects, $Z \in \mathbb{R}^{n\times b}$ is the design matrix of the random effects, $\boldsymbol{u} \in \mathbb{R}^b$ is the vector of random effects, $\boldsymbol{\xi} \in \mathbb{R}^n$ is the vector of spatially dependent random error, and $\boldsymbol{\eta} \in \mathbb{R}^n$ is the zero mean random vector of pair-wise independent errors. In particular, $\boldsymbol{\xi}$ is modeled as an anisotropic separable bivariate AR process commonly referred to as AR1×AR1. This mixed-effect formulation also allows to estimate all variance parameters using REML, [30]. Important in NNA and AR models is the assumption that autocorrelation between the measurements decays exponentially with distance, which is a property of AR processes. In applications of this model the user needs to inspect the variogram of the estimated residuals and determine whether the error structure is devoid of spatial effect, if not, a new iteration should be applied. The model also allows for the addition of terms to estimate other sources of spatial variation known to the user.

We estimate and remove the field-induced spatial variation with the state of the art spatial-effect removal method introduced in [52], and implemented in the R-package SpATS [53]. This method models the vector $\boldsymbol{y} \in \mathbb{R}^{720\times 1}$ of raw measured canopies from 720 plots arranged in 60 rows and 12 columns and featuring 180 seedlots as follows:

$$\boldsymbol{y} = f(\boldsymbol{w}, \boldsymbol{v}) + X_s\boldsymbol{\tau}_s + Z_c\boldsymbol{u}_c + Z_r\boldsymbol{u}_r + Z_b\boldsymbol{u}_b + \boldsymbol{\varepsilon}, \tag{2.1}$$

where $\boldsymbol{y} \in \mathbb{R}^{720\times 1}$ is the vector of raw measured canopies, $f(\boldsymbol{w}, \boldsymbol{v})$ is the fitted two dimensional surface, $X_s \in 720 \times 180$ is the design matrix assigning measurements to seedlots interpreted as fixed effects, $Z_c \in \mathbb{R}^{720\times 12}$ is encoding the column information, $Z_r \in \mathbb{R}^{720\times 60}$ is encoding the row information, and $Z_b \in \mathbb{R}^{720\times 4}$ encodes the block information. All the $Z$ matrices refer to random effects and the $X$ matrix corresponds to the fixed effect.

The package provides an estimate of the smooth field variation, models row, columns, and block effects and provides the Best Linear Unbiased Estimate (BLUE) of the mean seedlot canopy size. The presence of an user-friendly implementation, and good convergence behavior motivated our choice of this method and software for this project. The typical output of the SpATS package is illustrated in Figure 2.2. In Section 5 of [52] the proposed model, SpATS, is benchmarked against AR1×AR1, showing that the performance of the two is similar, with a net improvement in terms of convergence demonstrated by SpATS.

The spatial variation is modeled as a two dimensional smooth surface using $p$-splines and, thanks to the mixed-model representation of p-splines, the spatial variation and the genotype and block effects can be modeled simultaneously. In contrast to previous smoothing methods, $p$-splines allow to model two-dimensional smooth trends, both globally and locally without the need to include spatially-correlated components. Variances for the residual and the noise in the random effects are found through REML.

We apply the spatial-effect removal to the raw canopy data obtained from all available orthophotos. We retrieve the BLUE of the seedlot canopy, which corresponds to the interpretation of the seedlot as a fixed effect. Therefore, after this correction step, we obtain a vector of size 180 per field and measurement, containing one estimate of the canopy mean per seedlot.

## 2.3 Specialized methods and algorithms

Given the large scale of the project and the level of precision to which we observed the canopies, the development of ad hoc techniques was necessary.

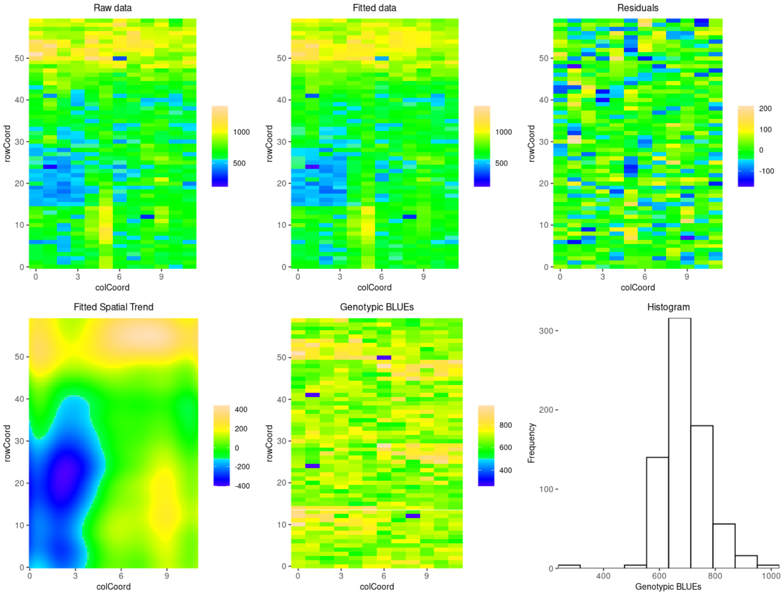


Figure 2.2: Spatial plot produced by the package SpaTS as an illustration of spatial effect removal.

In this section we detail our methodology for measuring the plant vitality, with particular emphasis on the procedure developed for the field trials.

### 2.3.1 Plot localization

Due to the relatively large scale of each trial (180 seedlots with 4 repetitions) and limited human resources, the chosen planting technique, which involved manual placing of the seed tubers from a moving vehicle, resulted in the trial fields that did not exhibit the usual regular plot structure. A regular structure, with the easily identifiable rows and columns of plots, aids the subsequent algorithmic plot identification and demarcation in the remote sensing data, e.g., drone images. In the absence of plot regularity, one has to isolate all plots in the drone images by relying on some kind of ground marks or image features.

In our case, six tubers were placed along each of the four neighboring ridges of the plot. Then, a small gap was left along each of the four ridges before planting the tubers of the next plot, which resulted in inter-plot gaps. These gaps become visible when all plants have emerged but are not yet too large, i.e., their canopies do not bridge and cover the gaps.

We have exploited these peculiarities by detecting the plot boundaries in a single suitable image and transforming (overlaying) these plot boundaries onto other images of the season, taken on both the preceding the subsequent dates. The main steps of the applied plot-detection algorithm are:

1. From the provided row-column plot labeling and schematics, the expected number $N$ of plots along the ridges of the trial fields is identified
2. We identify the field image best suited for gaps detection. Such an image is usually found towards the end of the canopy growth season, where canopies inside a plot are touching, but have not yet grown as to bridge the inter-plot gaps
3. The beginning and the end of the trial field along each ridge is interactively determined in the selected image
4. The expected number of $N - 1$ inter-plot gaps is automatically detected in the images along each ridge. To this end:
   (a) The field image is binarized using a strict green filter such that decidedly green regions will be white and soil will be black in the resulting image. Since the purpose is to find gaps between lots the use of a strict filter is not detrimental in this case
   (b) The morphological operation of closure is applied to the binarized image to discard possible noise left over from the binarization procedure
   (c) Each ridge is uniformly divided into $N$ sub-intervals
   (d) Image intensity is extracted along three lines parallel and in the neighborhood of the mid-ridge line
   (e) The regions for which the intensity along all three lines is zero (black, i.e. soil in our binarization) are considered the inter-plot gaps and the image coordinates of the midpoint of these regions is saved for further processing
5. The detected plot polygons are displayed and inspected for eventual remaining inaccuracies and distortions and the wrongly identified plot boundary points are corrected interactively
6. For each plot a set of image coordinates of the plot polygonal boundary is saved

Examples of detected inter-plot gaps along the ridges can be found in Figure 2.3. For an example seedlot, the four plots detected in a trial field are shown in Figure 2.4.

### 2.3.2 Time alignment and distortion correction

The stitched RGB orthophotographic images of the trial fields obtained at several dates during the growth season are not spatially aligned between the dates, i.e., the same pixel coordinates in two images do not correspond to the same locations on the ground. In order to overlay the polygons (plot boundaries) found on a given date on the images taken on other dates, we look for the transformations between the

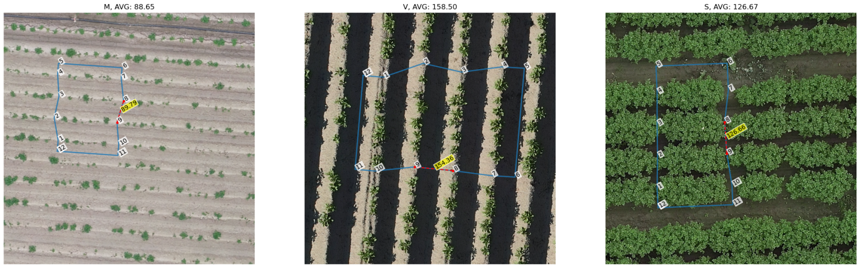


Figure 2.3: Examples of polygon boundary vertices found in different fields at different times throughout the season with vertex labels. Numbers inside yellow rectangles are the distances between the two vertices in pixels. Image titles show the field average of the inter-ridge distance in pixels, which is subsequently used for the conversion of the canopy area measurements to cm$^2$.

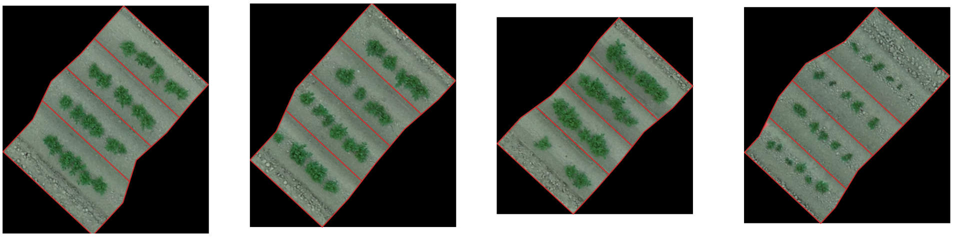


Figure 2.4: The four plots of the same batch detected in the field on a given date. Also shown are the three inter-ridge boundaries within the plots.

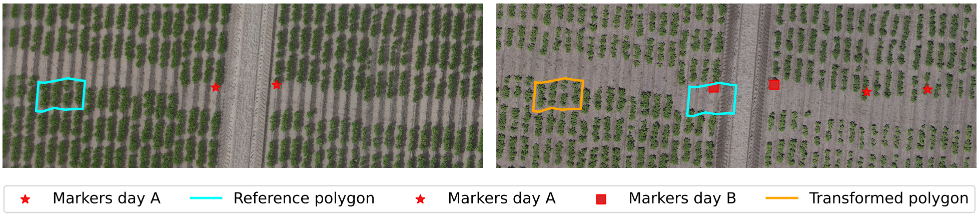


Figure 2.5: This figure illustrates the need for alignment. The markers and polygon found in day A (left panel) are not in the right spot in the coordinate system of day B (right panel), but using the day specific marker coordinates we can correctly transform day A's polygons to the coordinates of day B (orange polygon in the right panel) without needing to recompute them.

reference frame of the image selected for plot detection and all other images of the growth season. We call this procedure the *time alignment* of the orthophotographs. While, in general, this transformation can be rather complicated, in the majority of the cases, an affine transformation turned out to be completely sufficient, as was confirmed by the visual inspection of the transformed superimposed polygonal plot boundaries.

Time alignment relies on the presence of time-invariant features, visible in all images, with the same physical position on the ground. In the 2019 trials we had to rely on natural time-invariant features, such as the connectors of irrigation pipes. In the 2020 and 2021 trials we had artificial, square, 10 cm × 10 cm, red-colored markers installed in the fields that were visible in all RGB images and we could use as reference points. Unfortunately, these markers turned out to be invisible in the multi-spectral images, which prevented us from time-aligning and using these valuable data.

The detection of the red markers is aided by the pre-processing of the images, where bright red pixels are identified and grouped so that their position can be highlighted on an interactive image plot. The user is shown the image in question with dots in contrasting color plotted at the positions of the algorithmically detected markers. The user has to zoom in on the relevant portions of the field and manually select the geometric middle point of the markers. This is an important step, since the accuracy of the marker locations on the orthophoto directly influences the accuracy of the transformation.

To understand the simple mathematics behind the recovery of the affine transformation, let there be $N$ markers with their (pixel) coordinates $(x_i, y_i)$, $i = 1, \dots, N$, on the reference date stored in the vector $\mathbf{p} \in \mathbb{R}^{2N}$ as $\mathbf{p}^T = [x_1, y_1, \dots, x_N, y_N]$. Suppose that the same $N$ markers have the coordinates $(\tilde{x}_i, \tilde{y}_i)$, $i = 1, \dots, N$ in another orthophoto taken on another date. Since a two-dimensional affine transformation is defined by just four numbers, two or more markers with known (same) physical positions in both orthophotos are sufficient. Although, the more markers one has, the more robust is the recovery of the transformation against the noise. With $N$ markers the four transformation elements stored in the vector $\mathbf{t} \in \mathbb{R}^4$ can be recovered by solving the following linear algebraic problem:

$$\begin{bmatrix} 1 & 0 & x_1 & -y_1 \\ 0 & 1 & y_1 & x_1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & x_N & -y_N \\ 0 & 1 & y_N & x_N \end{bmatrix} \begin{bmatrix} t_1 \\ t_2 \\ t_3 \\ t_4 \end{bmatrix} = \begin{bmatrix} \tilde{x}_1 \\ \tilde{y}_1 \\ \vdots \\ \tilde{x}_N \\ \tilde{y}_N \end{bmatrix}. \tag{2.2}$$

Given a sufficient number of markers, there exists the unique least-squares solution $\hat{\mathbf{t}}^T = [\hat{t}_1, \hat{t}_2, \hat{t}_3, \hat{t}_4]$ to this problem.

The obtained least-squares solution of the problem (2.2) can be used to determine the location of each vertex of the polygonal plot boundary in the second orthophoto.

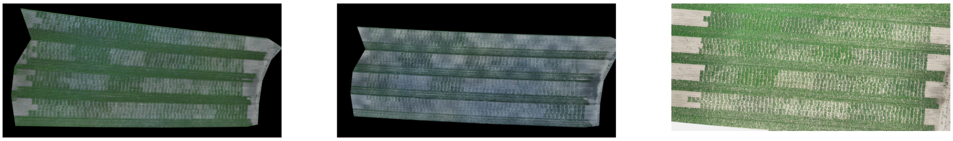


Figure 2.6: Original distorted orthophotograph of the Montfrin field taken in 2019 (left). The same image with distortion corrected by warping to achieve approximately the same inter-ridge distance across the field (middle). High-quality orthophoto of the Montfrin field taken in 2021 (right) confirms that the distortion in 2019 was and artifact of image stitching.

Let the (pixel) vertex coordinates on the reference date be $(x, y)$, then the pixel coordinates $(\tilde{x}, \tilde{y})$ of this vertex in the orthophoto from another date can be obtained as:

$$\begin{bmatrix} \tilde{x} \\ \tilde{y} \end{bmatrix} = \begin{bmatrix} \hat{t}_1 \\ \hat{t}_2 \end{bmatrix} + \begin{bmatrix} \hat{t}_2 & -\hat{t}_3 \\ \hat{t}_3 & \hat{t}_2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}. \tag{2.3}$$

While the above affine transformation was sufficient for the time alignment of all images in 2020, in 2019 and in one field in 2021, this transformation did not produce the desired results. Namely, applying (2.3) with the parameters obtained by solving (2.2) in the least-squares sense, yielded the transformed markers of the first field that were still not aligned with the ground markers of the second field. This means that, apart from the shift, rotation and anisotropic (but uniform) scaling, an additional non-linear distortion is present in one or both images. Such a distortion could be caused, for example, by a particularly nonuniform flight path of the drone that was not properly reconstructed during the stitching procedure. In the case where the affine transformation was insufficient, the polygons were transformed using the radial-basis interpolator function `rbf()` from the Python `scipy` package.

Having taken care of all relative distortions between the images taken on different dates, we have noticed an additional strong distortion in all images of the Montfrin test field in 2019, see Figure 2.6, which made the the presumably parallel ridges diverge towards one side of the field. The fact that it was a distortion of the image rather than the natural shape of the field is obvious from the high-quality orthophotograph of the same field taken in 2021 (Figure 2.6, right). Since this distortion has consequences for both the estimation of the green area and the spatial correction, the original orthophoto was re-processed. The field image was warped with the function `warpPerspective()` from the Python `opencv` package, so that the inter-ridge distance in the "left" part of the image, as measured with the help of selected polygonal vertices, became equal to the inter-ridge distance measured on the "right". The same transformation was used to warp the polygonal plot boundaries. No such distortions were observed or had to be corrected in other fields and years.

### 2.3.3 Leaf canopy segmentation and estimation

To measure the canopy area within the polygonal plot boundary, the image pixels are segmented into two disjoint sets: pixels of the canopy and pixels of the surrounding soil. Then, the canopy pixels are counted and the result converted to the $\mathrm{cm}^2$ units.

While a human operator is usually very successful in segmenting an image, a fully automated segmentation procedure that would work in all circumstances is not available. The present dataset featured a variety of illumination and moisture conditions, both of which affect the color of pixels. Also, leaf canopy colors have systematic differences between genotypes, ranging from light green to almost purple. Therefore, every orthophoto had to be processed individually, resulting in different segmentation filters with date- and field-specific parameters. In all cases, the quality of segmentation has been confirmed by visual inspection of randomly selected plots of each genotype.

The green segmentation procedure consists of the following steps:

1. Convert the image to HSV format
2. Find date-specific range of the Hue channel for the canopy. This is done by analyzing the Hue channel histogram, which shows two peaks, canopy and soil, with the canopy peak growing in time. The Otsu method finds a good first guess of the lower Hue boundary. This guess can be refined by inspection of sample plots. Set all the pixels outside the Hue range to 'black'
3. Equalize Saturation and Value channels
4. Filter the Value channel by allowing only the pixels with the value below a threshold and setting all other pixels to 'black'
5. Obtain a grayscale image by applying a weighted combination (weights are manually adjusted to achieve the best segmentation) of some standard agricultural 'green' filters. The Hue Index (HI) [45], Excess Green (EXG) [19], and Brightness Index (BI) [25]:

$$\mathrm{BI} = \sqrt{\frac{r^2+g^2}{2}} \tag{2.4}$$

$$\mathrm{EXG} = 2g-(r+b) \tag{2.5}$$

$$\mathrm{HI} = \frac{2r-g-b}{g-b} \tag{2.6}$$

6. Normalize the image, set all pixels below the threshold in column "thresh." in Table 2.5 to 'black', and binarize the image
7. Remove 'salt and pepper' noise by applying median blur with 3 × 3 kernel
8. White pixels – canopy, black pixels – soil

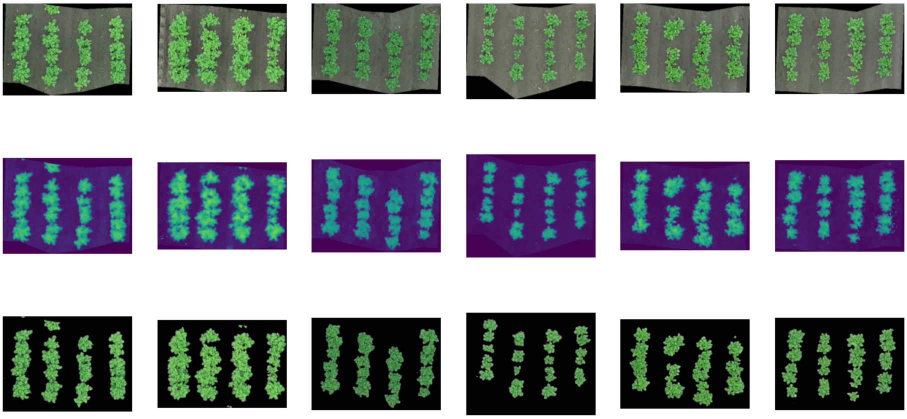


Figure 2.7: This image shows an example of plot segmentation for each variety (from left to right: Festien, Colomba, Seresta, Challenger, Sagitta, Innovator). The first row contains the original RGB image cropped to the plot boundary, the middle row shows the average of the RGB vegetation indices according to the day-specific weights, the bottom row shows the segmented images where all soil pixels are set to black.

The parameters of the above segmentation procedures are provided in Table 2.5 for each field and date.

After segmentation, the mean canopy area $S_{\mathrm{px}}$ (in pixels) over each plot is determined by summing all white pixels within the geometrical boundaries of the plot and dividing by 24 – the number of plants in each plot. To convert a canopy area in pixels to its area in cm$^2$, we use the fact that the distance $d_{\mathrm{cm}}$ between the ridges in the field is determined by the planting device and is $d_{\mathrm{cm}} = 75$ cm in the Veenklooster (V) and Kollumerwaard-SPNA (S) fields, and $d_{\mathrm{cm}} = 74$ cm in the Montfrin (M) field.

To find the pixel-to-cm conversion factor, we compute the average pixel distance $d_{\mathrm{px}}$ between the adjacent ridges in the field for a specific date (see Figure 2.3). Then, the area $S_1$ of a single pixel in cm$^2$ is given by:

$$S_1 = \left(\frac{d_{\mathrm{cm}}}{d_{\mathrm{px}}}\right)^2. \tag{2.7}$$

Thus, the canopy area $S$ in cm$^2$ is obtained from the canopy area $S_{\mathrm{px}}$ in pixels as:

$$S = S_1 S_{\mathrm{px}}. \tag{2.8}$$

### 2.3.4 Climate room data

The measurement of vitality from climate room images poses different challenges with respect to the experimental field setup, thus a different algorithm is necessary

| field/date | H range | S range | S,V eq. | V filter | (BI,HI,EXG) | thresh. | blur |
|---|---|---|---|---|---|---|---|
| M-19-04-10 | $[30, 80]$ | $[0, 255]$ | yes | $[0, 100]$ | 0.8, 0.8, 1.0 | 0.5 | yes |
| M-19-04-19 | $[25, 100]$ | $[0, 255]$ | no | $[0, 110]$ | 0.1, 0.3, 0.6 | 0.25 | yes |
| M-19-04-26 | $[25, 80]$ | $[0, 255]$ | no | no | 0.0, 0.0, 1.0 | 0.35 | no |
| V-19-05-24 | $[34, 180]$ | $[52, 255]$ | – | – | 0.0, 0.0, 1.0 | 0.2 | yes |
| V-19-05-29 | $[34, 180]$ | $[52, 255]$ | – | – | 0.0, 0.0, 1.0 | 0.2 | yes |
| V-19-06-07 | $[54, 180]$ | $[52, 255]$ | – | – | 0.0, 0.0, 1.0 | 0.2 | yes |
| S-19-06-07 | $[36, 180]$ | $[18, 255]$ | – | – | 0.0, 0.0, 1.0 | 0.2 | yes |
| S-19-06-19 | $[40, 180]$ | $[41, 255]$ | – | – | 0.0, 0.0, 1.0 | 0.2 | yes |
| M-20-04-13 | $[31, 180]$ | $[30, 255]$ | – | – | 0.2, 0.2, 0.6 | 0.45 | no |
| M-20-04-16 | $[32, 180]$ | $[30, 255]$ | – | – | 0.2, 0.2, 0.6 | 0.45 | yes |
| M-20-04-18 | $[34, 180]$ | $[52, 255]$ | – | – | 0.2, 0.2, 0.6 | 0.45 | yes |
| M-20-04-22 | $[38, 180]$ | $[52, 255]$ | – | – | 0.2, 0.2, 0.6 | 0.45 | yes |
| M-20-04-25 | $[36, 180]$ | $[47, 255]$ | – | – | 0.2, 0.2, 0.6 | 0.45 | yes |
| V-20-05-27 | $[36, 180]$ | $[33, 255]$ | – | – | 0.0, 0.0, 1.0 | 0.2 | yes |
| V-20-05-30 | $[27, 80]$ | $[60, 255]$ | – | – | 0.0, 0.0, 1.0 | 0.2 | yes |
| V-20-06-07 | $[44, 180]$ | $[41, 255]$ | – | – | 0.0, 0.0, 1.0 | 0.2 | yes |
| V-20-06-10 | $[50, 180]$ | $[50, 255]$ | – | – | 0.0, 0.0, 1.0 | 0.2 | yes |
| V-20-06-12 | $[50, 180]$ | $[55, 255]$ | – | – | 0.0, 0.0, 1.0 | 0.2 | yes |
| S-20-06-03 | $[44, 120]$ | $[80, 255]$ | – | – | 0.0, 0.0, 1.0 | 0.2 | yes |
| S-20-06-10 | $[35, 120]$ | $[70, 255]$ | – | – | 0.0, 0.0, 1.0 | 0.2 | yes |
| S-20-06-12 | $[35, 120]$ | $[75, 255]$ | – | – | 0.0, 0.0, 1.0 | 0.25 | yes |
| S-20-06-15 | $[35, 120]$ | $[75, 255]$ | – | – | 0.0, 0.0, 1.0 | 0.15 | yes |
| S-20-06-19 | $[40, 120]$ | $[52, 255]$ | – | – | 0.0, 0.0, 1.0 | 0.2 | yes |
| M-21-04-02 | $[35, 75]$ | $[0, 255]$ | – | – | 0.2, 0.2, 0.6 | 0.45 | yes |
| M-21-04-07 | $[26, 75]$ | $[0, 255]$ | – | – | 0.2, 0.2, 0.6 | 0.45 | yes |
| M-21-04-08 | $[28, 75]$ | $[0, 255]$ | – | – | 0.2, 0.2, 0.6 | 0.45 | yes |
| M-21-04-16 | $[35, 75]$ | $[0, 255]$ | – | – | 0.2, 0.2, 0.6 | 0.45 | yes |
| M-21-04-19 | $[35, 100]$ | $[0, 255]$ | – | – | 0.2, 0.2, 0.6 | 0.45 | yes |
| M-21-04-22 | $[35, 90]$ | $[0, 255]$ | – | – | 0.2, 0.2, 0.6 | 0.45 | yes |
| M-21-04-26 | $[30, 75]$ | $[0, 255]$ | – | – | 0.2, 0.2, 0.6 | 0.45 | yes |
| M-21-05-02 | $[36, 90]$ | $[0, 255]$ | – | – | 0.2, 0.2, 0.6 | 0.45 | yes |
| M-21-05-04 | $[38, 100]$ | $[0, 255]$ | – | – | 0.2, 0.2, 0.6 | 0.45 | yes |
| M-21-05-08 | $[36, 90]$ | $[0, 255]$ | – | – | 0.2, 0.2, 0.6 | 0.45 | yes |
| M-21-05-11 | $[38, 80]$ | $[0, 255]$ | – | – | 0.2, 0.2, 0.6 | 0.45 | yes |
| M-21-05-14 | $[38, 75]$ | $[0, 255]$ | – | – | 0.2, 0.2, 0.6 | 0.45 | yes |
| M-21-05-19 | $[40, 80]$ | $[0, 255]$ | – | – | 0.2, 0.2, 0.6 | 0.45 | yes |
| V-21-05-20 | $[26, 50]$ | $[0, 255]$ | – | – | 0.2, 0.2, 0.6 | 0.45 | yes |
| V-21-05-24 | $[34, 75]$ | $[0, 255]$ | – | – | 0.2, 0.2, 0.6 | 0.45 | yes |
| V-21-05-28 | $[26, 75]$ | $[0, 255]$ | – | – | 0.2, 0.2, 0.6 | 0.45 | yes |
| V-21-05-31 | $[26, 75]$ | $[0, 255]$ | – | – | 0.2, 0.2, 0.6 | 0.45 | yes |
| V-21-06-04 | $[34, 75]$ | $[0, 255]$ | – | – | 0.2, 0.2, 0.6 | 0.45 | yes |
| V-21-06-07 | $[26, 75]$ | $[0, 255]$ | – | – | 0.2, 0.2, 0.6 | 0.45 | yes |
| V-21-06-11 | $[26, 75]$ | $[0, 255]$ | – | – | 0.2, 0.2, 0.6 | 0.45 | yes |
| V-21-06-18 | $[34, 75]$ | $[0, 255]$ | – | – | 0.2, 0.2, 0.6 | 0.45 | yes |
| V-21-06-23 | $[34, 60]$ | $[0, 255]$ | – | – | 0.2, 0.2, 0.6 | 0.45 | yes |
| S-21-06-18 | $[31, 50]$ | $[0, 255]$ | – | – | 0.2, 0.2, 0.6 | 0.45 | yes |
| S-21-06-23 | $[30, 50]$ | $[0, 255]$ | – | – | 0.2, 0.2, 0.6 | 0.45 | yes |
| S-21-06-28 | $[24, 40]$ | $[0, 255]$ | – | – | 0.2, 0.2, 0.6 | 0.45 | yes |
| S-21-07-03 | $[28, 62]$ | $[0, 255]$ | – | – | 0.2, 0.2, 0.6 | 0.45 | yes |
| S-21-07-05 | $[34, 60]$ | $[0, 255]$ | – | – | 0.2, 0.2, 0.6 | 0.45 | yes |
| S-21-07-08 | $[31, 62]$ | $[0, 255]$ | – | – | 0.2, 0.2, 0.6 | 0.45 | yes |
| S-21-07-14 | $[36, 75]$ | $[0, 255]$ | – | – | 0.2, 0.2, 0.6 | 0.45 | yes |
| S-21-07-16 | $[34, 62]$ | $[0, 255]$ | – | – | 0.2, 0.2, 0.6 | 0.45 | yes |

Table 2.5: Field- and date-specific parameters for canopy segmentation in orthophoto's.

to identify the seedlots.

Pots in the climate rooms are also distributed according to a RCBD on two tables per condition. In each genotype block a seedlot is represented by a box of $2 \times 2$ pots as visible in Figure 2.8. The pictures are taken from a rolling camera scaffolding taking care that consecutive pictures overlap by a third, since we originally thought this would be sufficient for the stitching of a table orthophoto. Unfortunately, the distance between the camera and the table is too small to obtain satisfactory results with the available open-source software WebODM. In the absence of an orthophoto, our approach to the detection and association of the seedlot box needed substantial modifications with respect to the field algorithm. In this case we exploited the fact that the table sides, as well as the labels, are found more or less in the same region of each image and that the pots are placed on trays. The pot recognition steps are as follows:

- With the Python package `pytesseract` and user confirmation when needed, we detect the labels at the center of the table
- based on the label number we can refer to the table scheme in order to know which seedlots are present in the image and the images of which seedlot containers are complete; note that in Figure 2.8 the seedlots in row 55 and 48 are only partially visible, and will have to be extracted from another image
- based on the horizontal coordinates of the label strip and the vertical coordinates of each single label we plot a fixed lattice, in red in Figure 2.8, bounding the boxes of the fully visible seedlots, and calculate the pot centroids, blue dots in Figure 2.8
- the calculated lattice is displayed over an interactive image and the user is given the opportunity to adjust the lattice by shifting it up or down
- upon receiving the confirmation from the user, the coordinates of the bounding boxes and the pots' centroids are stored in a `csv` file with the corresponding seedlot name and the image file name.

Once an association between pixel coordinates and seedlots is established, one can proceed with the canopy segmentation in order to measure the canopy size. However it was already possible to see at that stage that the canopy extraction will not be very accurate. For example, in the bottom panel of Figure 2.8, one can see that plants do not remain within the pot boundaries and, starting with the DAP 36, part of this seedlot is occupying the neighboring box.

### 2.3.5 HSI segmentation

Hyperspectral imaging (HSI) captures information regarding both the composition and the distribution of materials in a sample, with the material composition indirectly characterized by the local infrared spectrum. Thus, an HSI image is a three-

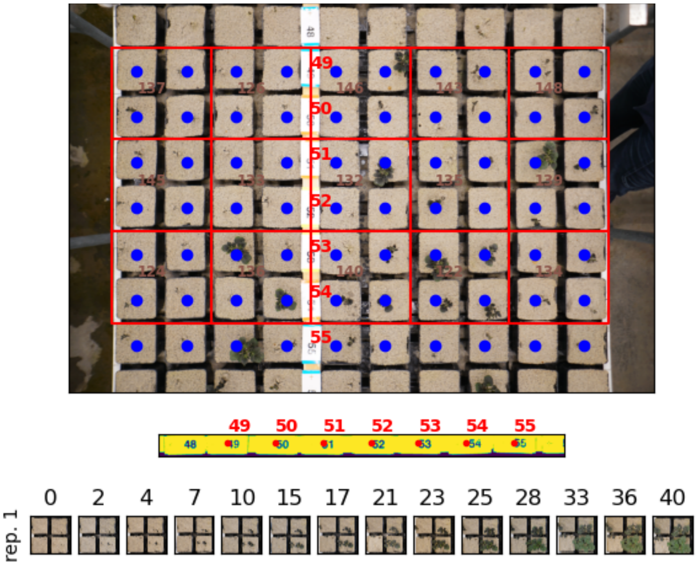


Figure 2.8: The figure illustrates the pot-detection algorithm in the climate room. Each 2 × 2 group of pots contains the same seedlot. In this final version of the experimental set up the pots sit on plastic trays in order to have a fixed distance and colored numbered tags in the middle of the table indicate the row number to aid with seedlot association. The tags are read by the algorithm and the pot centroids and boundaries are found in each image, each pot is present in multiple consecutive images, so one image has to be selected from which to measure the canopy. We opt for the image in which the plant is seen from the top. In the sequence on the bottom of the figure we see the growth of one pot from 0 to 40 days after planting. Note that canopies develop beyond the pot boundaries.

dimensional array. In fact, a visual-spectrum RGB image, where each pixel is a weighted mixture of three colors, red, green and blue, and the image itself has two spatial dimensions, height and width, is also a three dimensional array. An HSI image features hundreds of 'color' channels, and covers a subset of the electromagnetic spectrum from near infrared to long-wave infrared wavelengths, so that each pixel contains a quasi-continuous spectral curve.

In our setup the images portray several slices of tubers of the same seedlot, see Figure 2.9 (upper-left image). Due to the expected biological differences that may possibly result in distinct infrared spectra, we divide each slice into two disjoint partitions: pith and cortex. The pith is the innermost part of the tuber, so also the innermost part of the slice, while the cortex in the part of the slice closest to the peel. The HSI seedlot data represent the average spectral signatures of the cortex and pith of all sampled slices of a seedlot, see Figure 2.9 (bottom plot).

Technically, the most involved stage in the extraction of the HSI seedlot data is the automated segmentation of the slices into the pith and cortex parts. The outer boundary of each slice is detected on the gray-scale image obtained by averaging over the first 50 spectral components. In this average image, the pixels corresponding to the table have a value of zero. Then, we detect the non-zero connected components whose area exceeds a threshold $\theta$ and compute their rectangular bounding boxes, see Figure 2.9 (upper-left image).

The cortex is a relatively thin outer layer of the tuber surrounding the inner core, the pith. In order to find such a boundary-conforming region for every slice, given the variation in size and the shape of each slice, we define a two-dimensional Poisson equation with the source at the detected centroid of the slice and the homogeneous Dirichlet boundary condition:

$$\begin{aligned} -\Delta u(\boldsymbol{x}) &= f(\boldsymbol{x}), \quad \boldsymbol{x} \in \Omega, \\ f(\boldsymbol{x}) &= \begin{cases} 1, & \boldsymbol{x} \in \Omega_{\mathrm{c}}, \\ 0, & \text{otherwise}, \end{cases} \\ u(\boldsymbol{x}) &= 0, \quad \boldsymbol{x} \in \partial\Omega, \end{aligned} \tag{2.9}$$

where $\Omega_{\mathrm{c}}$ is a single-pixel cell surrounding the centroid $\boldsymbol{x}_{\mathrm{c}}$, $\Omega \subset \mathbb{R}^2$ is the entire slice, and $\partial\Omega$ is its boundary. According to the Maximum Principle and the Positivity Theorem, [65], the solution $u(\boldsymbol{x})$ of the boundary-value problem (2.9) is positive inside $\Omega$, i.e., $u(\boldsymbol{x}) \geq 0$, $\boldsymbol{x} \in \Omega$, and, due to the imposed boundary condition, equals zero at the boundary $\partial\Omega$. Moreover, although it does not immediately follow from the mentioned theorems, it is easy to anticipate that the structure of the source will cause the solution, similarly to the Green's function of the Poisson equation, to reach its maximum at $\boldsymbol{x}_{\mathrm{c}}$. These facts give a natural way to segment the domain by defining the cortex as the subdomain of $\Omega$, bounded on the outside by the boundary $\partial\Omega$ and on the inside by the level curve $u(\boldsymbol{x}) = u_0$, $0 \leq u_0$.

To arrive at the solution $u(\boldsymbol{x})$, for each slice, we solve the boundary-value problem (2.9) numerically on the doubly-uniform Cartesian grid formed by the image pixels (pixels selected by the slice mask). We employ the standard second-order accurate Finite-Difference approximation of the Lapacian operator, which results in a sparse linear system of equations that is then solved by the direct sparse-LU solver implemented in the `spsolve()` method of the `scipy.sparse.linalg` library.

The main property of the obtained level curve $u(\boldsymbol{x}) = u_0$ is that it conforms to the outer boundary for the level $u_0 = 0$ and gradually changes its shape towards a shrinking circle around the centroid $\boldsymbol{x}_\mathrm{c}$ for higher positive values of $u_0$. For each slice, the level $u_0$ of the curve is chosen in such a way that the area of the cortex part constitutes 30% of the total slice area, see Figure 2.9 (upper-right image).

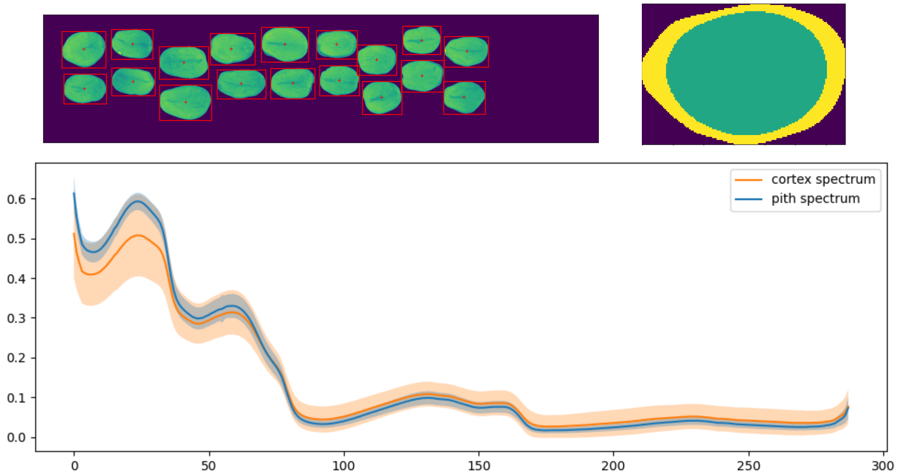


Figure 2.9: Extraction of the HSI spectral signatures. Top-left panel: spectrum-average image of the tray containing the seedlot tuber slices with the bounding boxes of individual slices. Top-right panel: the result of the segmentation of a single slice into its cortex (outer, yellow) and pith (inner, green) parts. Bottom panel: the spectra of all the cortex and pith pixels (shaded) and the corresponding average spectral signatures.

## 2.4 Choice of the vigor measure

There is no universally accepted measure of the plant vitality that is mentioned in the title of the project “Flight to Vitality”. In fact, although it may be a matter of semantics, one could argue that the term ‘vigor’ is more applicable in the present situation.

In any case, the size of the leaf canopy around the time of tuber bulking, determines the rate of tuber growth and the eventual yield of potato plants, which is the main trait of interest for potato farmers. Therefore, the sooner a plant can develop a large canopy the more vigorous or vital is this plant. From this point of view, it appears that considering the time evolution of the canopy could eventually lead to a measure of vigor. Unfortunately, due to sparsity and inconsistency of the imaging dates, a systematic analysis of the canopy time evolution could not be achieved in this project. In this circumstances, a well-chosen single-time measurement of the canopy size may also be considered an indicator of the plant vigor.

From the observational point of view, the relative difference in the vigor of potato plants can manifest itself in several ways. First, some seed tubers can sprout sooner than others. Second, the sprouts can grow at different rates, emerging above the ground at different times. Finally, the rates of growth of the leaf canopies can be different, so that canopies achieve different sizes at the moment when the plant blooms and produces the tubers. Hence, a large canopy at a certain point in time could be either due to an early-emergent but relatively slow-growing plant or due to a late-emergent and fast-growing plant. Nevertheless, a relatively large canopy somewhere in the middle of the growth season, after the early 'transients' related to the emergence time and the initial rate of growth have passed, is surely indicative of the plant vigor. Indeed, as long as a plant acquires a sufficiently large canopy at a physiologically useful point in time, e.g., around the time of tuber bulking, it can be considered a vigorous plant, and it does not really matter whether it is due to an early emergence time or a fast initial growth rate.

Figure 2.11, Figure 2.12, and Figure 2.13 show the summary of the BLUE canopy data for each of the six genotypes at the available time points for each year and field. As one can see, the growth pattern and the attained canopy size are different between the fields and years. This can be attributed to the varying weather conditions and soil types, caused mainly by the difference in geographic location and the different times of the year the seed tubers were planted in each field, see Table 2.1. Hence, a single *a priori* chosen day after planting cannot be used as a time point to take the vitality measurement. Therefore, our choice of the canopy measurement date is year- and field-specific and is guided by the data and the following basic principles:

- It should be possible to compare different seedlots within each genotype
- The plant-plant and seedlot-seedlot interactions should be avoided
- The growth has reached a phase in which the ordering of seedlots by their average canopy size is more stable

From this point of view, there are only one or two dates per year and field among the available data that can be used as the vitality measure. Indeed, in the early measurements not all plants have emerged yet, which does not allow for a fair comparison of the seedlots. In the late measurements, the canopies of rapidly growing genotypes start "merging" not only inside the plots, but also with those of the neighboring plots, thus introducing plant-plant and seedlot-seedlot interactions and make it impossible to properly estimate the canopy size from drone images.

As a measure of seedlot order stability, we have chosen to the Kendall $\tau$ on the vector of quantized canopy size, Figure 2.10. This means that we have labeled the canopy sizes as large, 1, average, 2, and small, 3, in each measurement date and calculated the Kendall $\tau$ on the vectors of these labels for all measurement dates of a field, pair-wise. The Kendall $\tau$ ranges from $-1$ to 1 and, in our case, quantifies the instability of the three canopy classes, i.e., if in two dates the labels of the seedlots have not changed, the Kendall $\tau$ will be equal to 1 and close to $-1$ if all labels have perfectly reversed.

From 47 to 50 DAP's all plants have emerged, the canopies are not yet overlapping, and the seedlot ranking as measured by the Kendall $\tau$ has stabilized making this a good time range to consider and reducing the choice to one or two dates per field in each year. From Figure 2.11, Figure 2.12, and Figure 2.13 it is clear that this corresponds to the end of the exponential growth period and precedes the period of 'saturation' in the expected sigmoid growth curves. In the year 2019, with only a few dates available, the choice was mostly dictated by the seedlot order stability. In 2020 and 2021 the 47-th DAP was chosen where available and in the test field V in 2020, the choice was again based on the date of maximal order stability. It should be mentioned that these final choices do not alter the conclusions of the subsequent association/regression studies in any significant way, since the data at these time points are very highly correlated, see Table 2.6.

In Figure 2.14 one can see how the choice of the DAP affects the (Pearson) correlation in the vigor parameters between the test fields – the quantity which is eventually displayed in Figure 2.15. It is clear that, due to the adverse weather conditions in France in 2021, no choice of DAP could significantly improve the correlations with the test field M in 2021.

| 2019 **M** | 52 |
|---|---|
| 45 | 96% |

| 2020 **M** | 47 | 50 |
|---|---|---|
| 43 | 96% | 96 % |
| 47 | - | 98% |

| 2021 **M** | 54 | 56 |
|---|---|---|
| 48 | 98% | 98 % |
| 54 | - | 99% |

| 2019 **S** | 48 |
|---|---|
| 36 | 90 % |

| 2020 **S** | 47 | 51 |
|---|---|---|
| 44 | 99% | 97 % |
| 47 | - | 99 % |

| 2021 **S** | 25 | 32 |
|---|---|---|
| 30 | 98% | 97 % |
| 25 | - | 97 % |

| 2019 **V** | 41 |
|---|---|
| 36 | 64% |

| 2020 **V** | 49 | 51 |
|---|---|---|
| 46 | 99% | 97% |
| 49 | - | 97% |

| 2021 **V** | 63 | 51 |
|---|---|---|
| 58 | 97% | 96% |
| 63 | - | 92% |

Table 2.6: Correlations (Pearson) of canopy size at neighboring DAPs in 2019 (left), in 2020 (center), and in 2021 (right). For all correlations holds $p < 0.001$.

Our final choices of the dates for the vitality measurement are highlighted in green

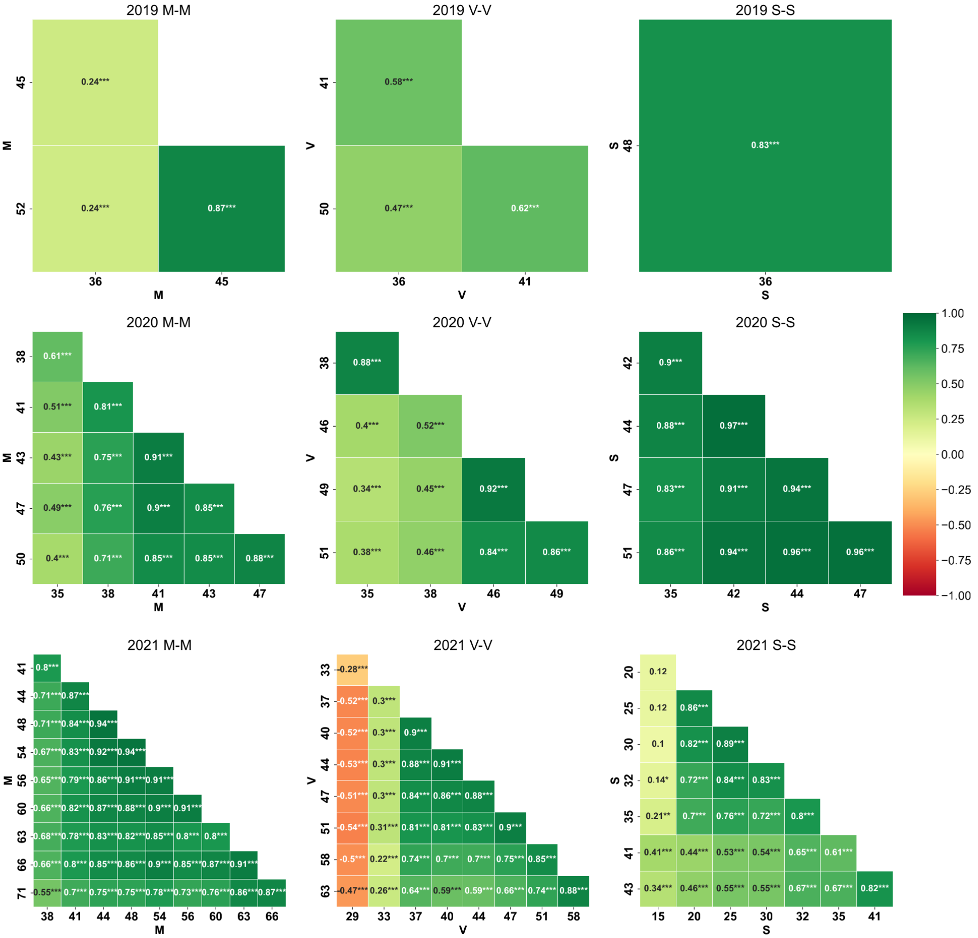


Figure 2.10: The Kendall $\tau$ coefficient is computed for each pair of measurements in one field. As we see the general trend is for the ordering to stabilize towards the end of the growth (bottom right of each correlation subplot). In this plot we show the stability measure with respect to a quantization of the vigor into three classes. We indicate the significance of the coefficient with stars, $*$ if $p < 0.05$, $**$ if $p < 0.01$, $***$ if $p < 0.001$. If no $*$ is present, the coefficient is not statistically significant.

along the $x$-axis of Figure 2.11, Figure 2.12, and Figure 2.13 and in Table 2.1 and are separately summarized in Table 2.7.

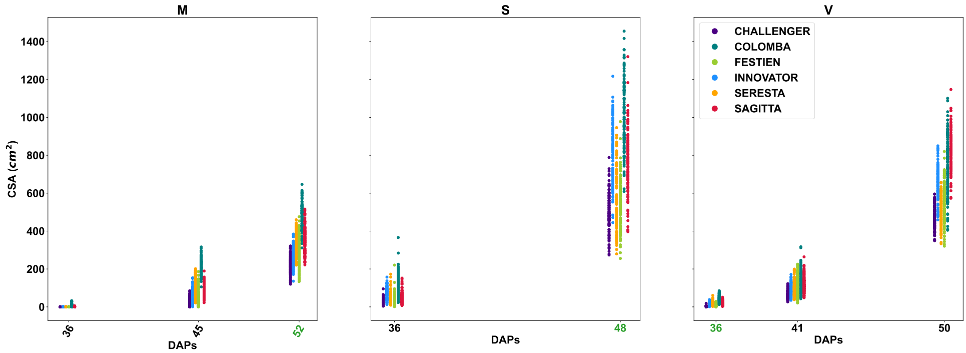


Figure 2.11: Summary of all available measurements in 2019 per genotype as scatterplots. The choice of the date for the vitality measurement is highlighted in green color on the $x$-axis.

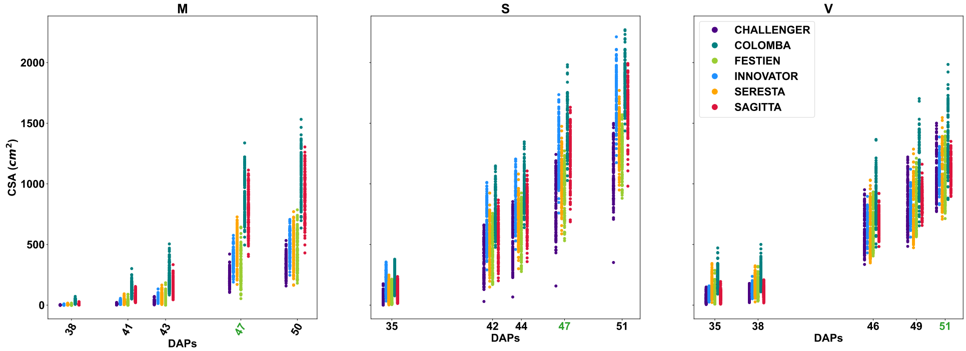


Figure 2.12: Summary of all available measurements in 2020 per genotype as scatterplots. The choice of the date for the vitality measurement is highlighted in green color on the $x$-axis.

## 2.5 Correlations in vigor across fields

After all the necessary steps for the vitality data extraction and removal of known effects, an exploratory analysis of the vitality data has been performed to ascertain the plausibility of the main project hypothesis that the vitality of a plant is determined, at least to a certain extent, by the seed tuber from which the plant has grown. The presence of such dependence could be inferred if the vitality of the plants produced

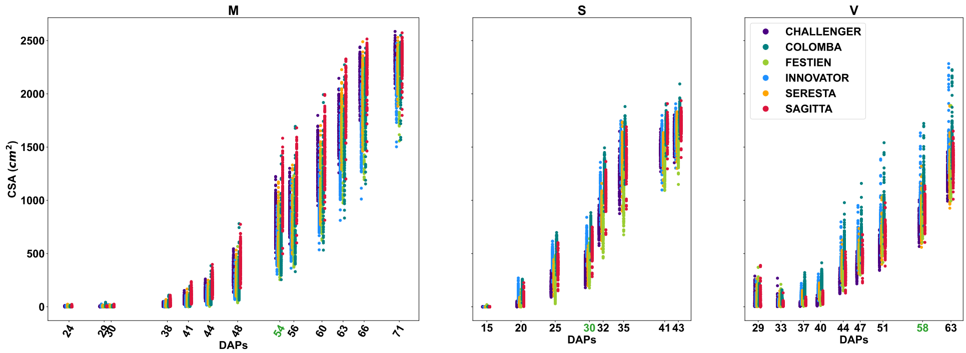


Figure 2.13: Summary of all available measurements in 2021 per genotype as scatterplots. The choice of the date for the vitality measurement is highlighted in green color on the $x$-axis.

| | M | S | V |
|---|---|---|---|
| 2019 | 52 | 48 | 36 |
| 2020 | 47 | 47 | 51 |
| 2021 | 54 | 30 | 58 |

Table 2.7: Measurement dates (in DAP's) per test field and year, when the canopy area is considered to be the seedlot vigor parameter.

by the seed tubers of a given seedlot, relative to other seedlots, is consistent for repetitions inside the field and across different fields.

The Pearson correlation coefficient of the vigor data provides an adequate measure of consistency. Correlation analysis can only be applied within a given year, where tubers of the same seedlot were planted in three different fields. As was mentioned above, while the same genotypes were tested in both years, the seed-tuber seedlots had a different production origin in each year and cannot be directly compared for consistency.

Correlations between the raw and spatially corrected (BLUE) vigor data across fields in each year are shown in Figure 2.15. One can see that both the raw and the BLUE data significantly correlate across the fields in all years, with the spatial effect removal leading to an increase in the correlation. It is also apparent that the field of Montfrin (M) in 2021 does not correlate with the other two experimental fields. We hypothesize that this lack of correlation is due to frost early in the growth season, which affected the early-emerging varieties (Challenger, Festien, and Seresta) and, on a finer scale, the early-emergent seedlots, thus resulting in a lack of correlation for the full field measurements. In general, the lack of correlation in other cases could be attributed to different weather and soil conditions, inconsistent application of herbicides, and

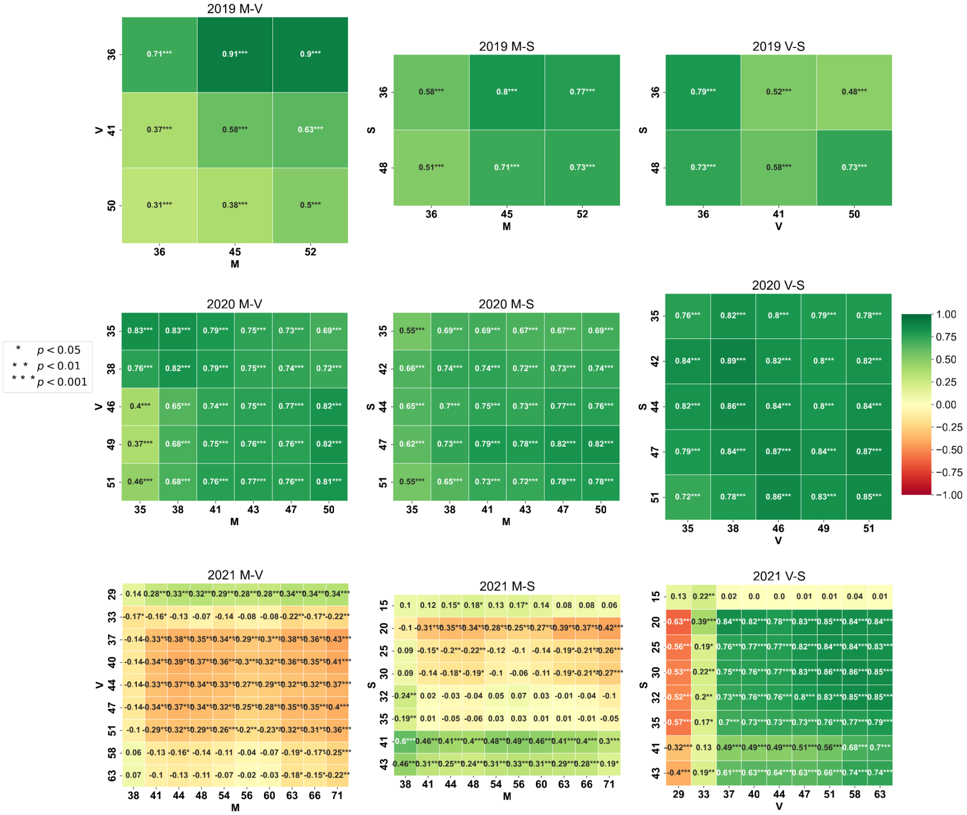
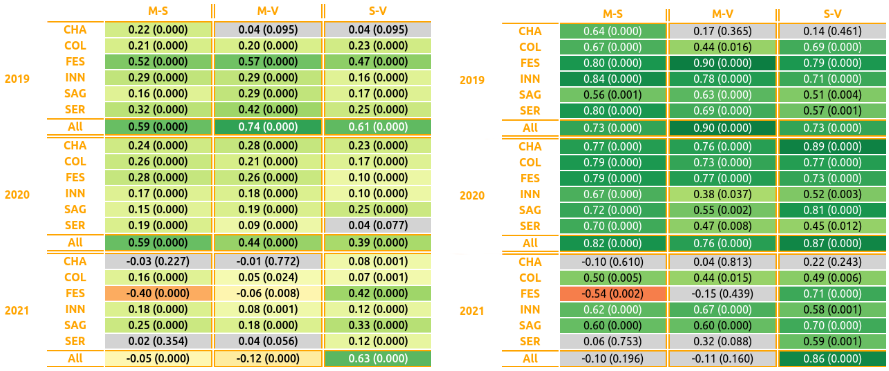


Figure 2.14: Pearson's correlation coefficient for the canopy sizes on different dates (DAP's) between the three test fields in each year.



| | | M-S | M-V | S-V |
|---|---|---|---|---|
| 2019 | CHA | 0.22 (0.000) | 0.04 (0.095) | 0.04 (0.095) |
| | COL | 0.21 (0.000) | 0.20 (0.000) | 0.23 (0.000) |
| | FES | 0.52 (0.000) | 0.57 (0.000) | 0.47 (0.000) |
| | INN | 0.29 (0.000) | 0.29 (0.000) | 0.16 (0.000) |
| | SAG | 0.16 (0.000) | 0.29 (0.000) | 0.17 (0.000) |
| | SER | 0.32 (0.000) | 0.42 (0.000) | 0.25 (0.000) |
| | All | 0.59 (0.000) | 0.74 (0.000) | 0.61 (0.000) |
| 2020 | CHA | 0.24 (0.000) | 0.28 (0.000) | 0.23 (0.000) |
| | COL | 0.26 (0.000) | 0.21 (0.000) | 0.17 (0.000) |
| | FES | 0.28 (0.000) | 0.26 (0.000) | 0.10 (0.000) |
| | INN | 0.17 (0.000) | 0.18 (0.000) | 0.10 (0.000) |
| | SAG | 0.15 (0.000) | 0.19 (0.000) | 0.25 (0.000) |
| | SER | 0.19 (0.000) | 0.09 (0.000) | 0.04 (0.077) |
| | All | 0.59 (0.000) | 0.44 (0.000) | 0.39 (0.000) |
| 2021 | CHA | -0.03 (0.227) | -0.01 (0.772) | 0.08 (0.001) |
| | COL | 0.16 (0.000) | 0.05 (0.024) | 0.07 (0.001) |
| | FES | -0.40 (0.000) | -0.06 (0.008) | 0.42 (0.000) |
| | INN | 0.18 (0.000) | 0.08 (0.001) | 0.12 (0.000) |
| | SAG | 0.25 (0.000) | 0.18 (0.000) | 0.33 (0.000) |
| | SER | 0.02 (0.354) | 0.04 (0.056) | 0.12 (0.000) |
| | All | -0.05 (0.000) | -0.12 (0.000) | 0.63 (0.000) |

| | | M-S | M-V | S-V |
|---|---|---|---|---|
| 2019 | CHA | 0.64 (0.000) | 0.17 (0.365) | 0.14 (0.461) |
| | COL | 0.67 (0.000) | 0.44 (0.016) | 0.69 (0.000) |
| | FES | 0.80 (0.000) | 0.90 (0.000) | 0.79 (0.000) |
| | INN | 0.84 (0.000) | 0.78 (0.000) | 0.71 (0.000) |
| | SAG | 0.56 (0.001) | 0.63 (0.000) | 0.51 (0.004) |
| | SER | 0.80 (0.000) | 0.69 (0.000) | 0.57 (0.001) |
| | All | 0.73 (0.000) | 0.90 (0.000) | 0.73 (0.000) |
| 2020 | CHA | 0.77 (0.000) | 0.76 (0.000) | 0.89 (0.000) |
| | COL | 0.79 (0.000) | 0.73 (0.000) | 0.77 (0.000) |
| | FES | 0.79 (0.000) | 0.77 (0.000) | 0.73 (0.000) |
| | INN | 0.67 (0.000) | 0.38 (0.037) | 0.52 (0.003) |
| | SAG | 0.72 (0.000) | 0.55 (0.002) | 0.81 (0.000) |
| | SER | 0.70 (0.000) | 0.47 (0.008) | 0.45 (0.012) |
| | All | 0.82 (0.000) | 0.76 (0.000) | 0.87 (0.000) |
| 2021 | CHA | -0.10 (0.610) | 0.04 (0.813) | 0.22 (0.243) |
| | COL | 0.50 (0.005) | 0.44 (0.015) | 0.49 (0.006) |
| | FES | -0.54 (0.002) | -0.15 (0.439) | 0.71 (0.000) |
| | INN | 0.62 (0.000) | 0.67 (0.000) | 0.58 (0.001) |
| | SAG | 0.60 (0.000) | 0.60 (0.000) | 0.70 (0.000) |
| | SER | 0.06 (0.753) | 0.32 (0.088) | 0.59 (0.001) |
| | All | -0.10 (0.196) | -0.11 (0.160) | 0.86 (0.000) |

Figure 2.15: Observed correlations (Pearson) in the canopy area on selected dates between the three fields (M, S, and V) in each year. The table on the left shows correlations for the raw measurements, the table on the right shows correlations for the spatially corrected measurements. The rows show the correlations for each variety as well as for all varieties together ('All'). Dark-green background highlights large significant positive correlations, red background indicates the significant negative correlation. The corresponding p-values are displayed between the brackets, insignificant correlations, i.e., with the p-value larger than 0.05, are displayed on a gray background.

the aging of seed tubers.

It is also clear that not all varieties show the same level of correlation between the test fields. One variety, namely, Festien, stands out as having the most consistent high correlations (70%-90%) when the M field in the year 2021 is excluded. The M-2021 field also shows the most significant negative correlation with respect to other fields for the Festien variety, which confirms our hypothesis that the early emergent and therefore normally vigorous seedlots were negatively affected by the cold temperatures early in the season. While high correlations are observed for other varieties as well, these are not as consistent. i.e., observed only between some of the test fields. Note also that only the correlations above 50% will result in useful predictions according to our metric, see (4.20).

# Chapter 3

# Overparameterized Multiple Linear Regression as Hyper-Curve Fitting

[1]

This chapter shows that the application of the fixed-effect multiple linear regression model to an overparameterized dataset is equivalent to fitting the data with a hyper-curve parameterized by a single scalar parameter. This equivalence allows for a predictor-focused approach, where each predictor is described by a function of the chosen parameter. It is proven that a linear model will produce exact predictions even in the presence of nonlinear dependencies that violate the model assumptions. Parameterization in terms of the dependent variable and the monomial basis in the predictor function space are applied here to both synthetic and experimental data. The hyper-curve approach is especially suited for the regularization of problems with noise in predictor variables and can be used to remove noisy and 'improper' predictors from the model.

## 3.1 Overparameterized regression

Many models encountered in practical applications of Machine Learning (ML) are overparameterized. Some are superficially so, e.g., a linear random-effect model [67] with many predictors that come from the same probability distribution, which by

[1] This chapter is based on E. Atza and N. Budko. Overparameterized multiple linear regression as hyper-curve fitting. *SIAM Journal on Mathematics of Data Science (Under review)*, 2024. URL `https://arxiv.org/abs/2404.07849`

itself is described by a few unknown parameters only. Others, like a linear fixed-effect model, also known as the Multiple Linear Regression (MLR) model, with fewer training samples than unknowns, may happen to be truly overparameterized. While there is no universally accepted definition of an overparameterized ML model, research in this area has recently uncovered several interesting phenomena, such as the double-dipping of the prediction error [34], and the so-called benign overfitting [11], [32].

The problems where the number of predictors is extremely large are common in many applications. Recent technological developments in chemistry, biology, medicine and agriculture have allowed for high-throughput data acquisition pipelines resulting in relatively large amounts of predictor variables compared to the feasible number of experiments in which a target phenotype or a trait is measured. For instance, in chemometric research, where one seeks to predict a physical or biological property from the infrared spectrum of the substance, the number of spectral components is in the order of thousands [43] [54]. In metabolomics, the biological traits are predicted from the relative abundance of small-molecule chemicals in a biological sample, the number of metabolites can reach tens of thousands [59], [66]. Similar numbers of predictors are used in the microbiome-based predictions [49]. Finally, in genomics, the number of Single-Nucleotide Polymorphisms (SNP's), that may potentially predict the phenotype of some living organism, can be as high as tens of millions [68], [56]. At the same time, the duration and costs of experiments aimed at measuring the dependent variables (phenotype, traits, etc.) are often much higher. Therefore, the number of such experiments, i.e., the number of training samples, is typically a fraction of the number of predictors (features) – hundreds or a few thousands at most [20], [1], [41], thus making the overparameterized nature of such problems inescapable.

Due to the enormous success of the Artificial Neural Networks (ANN's) in image classification, the current focus in the omics-related ML is on the application of these sophisticated nonlinear models to the existing and new data. However, truly deep ANN's are also overparameterized, which leads to learning problems when the number of training samples is low and require a large training/validation dataset to tune hyperparameters and avoid overfitting. Sometimes this problem is circumvented by creating an artificial augmented training dataset [16]. Therefore, it is not surprising that whenever the training is performed on small datasets, the quality of predictions obtained with an ANN is only marginally better, if at all, when compared to the predictions with the simple overparameterized MLR model [23], [33].

One could argue that the good results by the overparameterized MLR are also due to overfitting or a poor testing procedure (e.g. data leakage). While this certainly may be the case in some practical studies, generally, the fixed-effect MLR model with fewer training samples than predictors is well-understood on a theoretical level [63],[37]. The solution to the resulting underdetermined linear system, if it exists, is not unique. As the next best thing, one aims at recovering the minimum-norm least-squares (LS) solution, which always exists and is unique. The minimum-norm LS solution may, however, be sensitive to noise in the dependent variable. If the

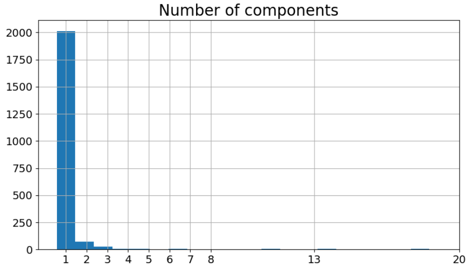


Figure 3.1: The figure shows an histogram of the cross-validated optimal number of PLS components for over 2000 models fitted on partitions of our data. It is clear to see that in the vast majority of cases the optimal number of components is chosen to be one.

level of noise in the dependent variable is known, then there is a regularized solution, which minimizes the error with respect to the noise-free minimum-norm LS solution. If, as it often happens in practice, the level of noise in the dependent variable is not known, then the level of regularization can either be estimated from the data with the cross-validation method or a similar technique, or has to be chosen subjectively. The case of the noise in predictor variables is much less understood, but there exist various errors-in-variables models [36].

In our research, we conducted an exhaustive association analysis on different chemical and biological datasets, most of which result in an underdetermined linear system. In a first instance, we chose to solve these by means of regularized partial least squares (PLS), where the dimension of the solution space is used for regularization. This study showed that in most cases the optimal dimension for the solution space is one, Figure 3.1. This result, which surprised us, motivated us to question ourselves about the need, true or perceived, to abandon linear models for more sophisticated non-linear ones.

The main focus of the present chapter is the adequacy, performance and optimization of overparameterized linear models. Specifically, the authors aimed at understanding the nature of overparameterized datasets, identifying the circumstances where the MLR model makes good or bad predictions of such datasets, improving the interpretability of regression results beyond the traditional feature weights, and increasing the prediction accuracy, simultaneously making the model more adequate, i.e., satisfying the linear assumptions. This has lead us to the column-centered reformulation of the MLR, where the (inverse) relation between each predictor variable and the dependent variable can to a large extent be analyzed independently of other predictor variables. We show that the predictions made by such an Inverse Regression (IR) model are identical to the predictions of the MLR model on a class of overparameterized datasets. Topologically this means that on such datasets the MLR model is not a hyper-plane as suggested by its mathematical form, but a hyper-curve, parameterizable by a single scalar parameter, which, for the sake of interpretability can be chosen as the dependent variable.

Due to its column-centered nature, the hyper-curve approach allows to filter out the predictors (features) that are either too noisy or do not satisfy the topological requirements of a linear model, which significantly improves the predictive power of the trained linear model, removes the features that may otherwise introduce the illusion of understanding [46], and suggests the subsets of predictors where a non-linear or a higher-dimensional-manifold model would be more adequate.

The chapter is structured as follows. In section 3.2 we define Fundamentally Over-parameterized (FOP) Datasets, MLR, PARametric hyper-CURve (PARCUR) and IR models, prove their equivalence and identify the condition for the exact prediction of a test dataset. In section 3.3 we study the behavior of the polynomial IR model applied to a dataset which contains both the polynomial predictors as well as predictors that have a non-functional relation to the dependent variable. We establish conditions under which such a dataset is a FOP dataset. In section 3.4 we consider noisy data and introduce a polynomial degree truncation regularization scheme that can handle the noise in both the dependent and predictor variables. In section 3.5 we propose a novel predictor removal algorithm that does not suffer from the ambiguities common to heuristic feature selection methods. In section 3.6 we apply the regularized IR model with predictor removal to the widely available experimental chemometric Yarn dataset [47], [60], and demonstrate the presence of both curve-like and higher-dimensional manifolds in this dataset. Finally, we present our conclusions and discuss the possible extensions of the PARCUR and IR models.

## 3.2 PARCUR and IR models

In this and the following section we focus on the mathematical properties of the model and all data is assumed to be exact. Data with additive noise will be considered in section 3.4. Given a sample $(y, x_1, \dots, x_p)$, the standard multiple regression model is a conditional parametric model whose goal is to recover the conditional distribution function $f_{Y|X=(x_1,\dots,x_p)}$ of the random variable $Y$ generating the observed $y$. In particular it expresses the conditional expectation of the dependent variable $Y$ as a function of the $p$ predictor variables $x_j$, $j = 1, \dots, p$, i.e.,

$$\mathbb{E}\left[y|x_1, \dots, x_p\right] = f(x_1, \dots, x_p), \quad f : \mathbb{R}^p \to \mathbb{R}, \tag{3.1}$$

where in the case of noiseless observations $\mathbb{E}\left[y|x_1, \dots, x_p\right] = \mathbb{E}\left[y\right] = y$. The Multiple Linear Regression (MLR) model, which assumes the function $f$ to be linear,

$$y = \sum_{j=1}^{p} \beta_j x_j, \tag{3.2}$$

obviously, belongs to this class of models. Topologically, the MLR equation (3.2) describes a $p$-dimensional linear object, a hyper-plane, in a $(p+1)$-dimensional space.

Now, consider the model of the form:

$$\begin{aligned} y &= x_0(s), \\ x_j &= x_j(s), \quad j = 1, \dots, p; \\ s &\in [a, b] \subset \mathbb{R}, \end{aligned} \tag{3.3}$$

which simply states that all data are considered to be the functions of some scalar parameter $s$. The equations (3.3), describe a parametric hyper-curve in a $(p+1)$-dimensional space, essentially, a one-dimensional object. We shall call this model the PARametric hyper-CURve (PARCUR) model.

Under a monotone transformation of variables, the PARCUR model is equivalent to the Inverse Regression (IR) model:

$$x_j = x_j(y), \quad j = 1, \dots, p. \tag{3.4}$$

In the inverse relation (3.4) the independent variables $x_j$, $j = 1, \dots, p$, are considered to be the functions of the dependent variable $y$. This model naturally emerges in the context of calibration problems [61]. It is also the most easily interpretable version of the PARCUR model as the functions $x_j(y)$ provide an insight into the change of each individual predictor variable as a function of the dependent variable $y$, if such a functional dependence exists.

Obviously, the general model (3.1) and the IR model (3.4) are completely equivalent only under very stringent constraints on the function $f$. Specifically, for a complete equivalence the function $f(x_1, \dots, x_p)$ as well as all individual functions $x_j(y)$ should be invertible. While an inverse of the MLR model (3.2) in the form (3.4) may exist on certain subsets of $\mathbb{R}^p$, a general nonlinear function $f$ in (3.1) is not invertible and neither is a general IR or PARCUR model. Yet, the predictive power of both the MLR and the PARCUR models trained on a finite training dataset of a certain general type appears to be the same even when they are not mutually invertible.

Our main results, expressed in theorem 3.3 and theorem 3.4, are the conditions on the exact prediction by the MLR model and the equivalence of the predictions made by the MLR and PARCUR models for what we call a Fundamentally OverParameterized (FOP) dataset. This equivalence allows to analyze the different types of column functions $x_j(s)$, $j = 1, \dots, p$, and establish the conditions on the existence of a FOP dataset for different types of predictor data.

In practice, any regression model is trained on a finite discrete dataset. An overparameterized dataset arises whenever the number $n$ of training samples is smaller than the number $p$ of parameters or predictors. This can happen simply due to the lack of experimental data and additional data can transform an overparameterized dataset into a well-defined or even an underparameterized one. However, in the present paper, we shall focus on the more fundamental case, where the simple addition of the training data does not change the overparameterized nature of the dataset.

**Definition 3.1.** The dataset

$$\mathcal{S} = \{(y_i, x_{i,1}, \dots, x_{i,p}),\ y_i, x_{i,j} \in \mathbb{R} |\ i = 1, \dots, m; j = 1, \dots, p\}$$

is a *fundamentally overparameterized* (in the linear sense) dataset of rank $q$, if, for any $m$, the data-matrix $S_m \in \mathbb{R}^{m\times(p+1)}$ with its rows from $\mathcal{S}$, i.e.,

$$S_m = [\boldsymbol{y}, X], \quad \boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix}, \quad X = \begin{bmatrix} x_{1,1} & \dots & x_{1,p} \\ \vdots & \ddots & \vdots \\ x_{m,1} & \dots & x_{m,p} \end{bmatrix},$$

has the property:

$$\operatorname{rank}(S_m) \le q \le p. \tag{3.5}$$

A *training* dataset is any fixed-size data matrix $S_n \in \mathbb{R}^{n\times(p+1)}$ with the rows from the FOP dataset $\mathcal{S}$. A training dataset is *complete* if $\operatorname{rank}(S_n) = q$. The complement dataset $\mathcal{S}_\mathrm{t} = \mathcal{S} \setminus \mathcal{S}_n$, is called the *test* dataset.

In practice we are dealing with arbitrary but fixed-size testing datasets as well. To summarize, the dependent-variable data from $S_n$ and $\mathcal{S}_\mathrm{t}$ will be stored in the data-vectors $\boldsymbol{y} \in \mathbb{R}^n$ and $\boldsymbol{y}_\mathrm{t} \in \mathbb{R}^m$, and the corresponding predictor-variable data will be stored in the data-matrices $X \in \mathbb{R}^{n\times p}$ and $X_\mathrm{t} \in \mathbb{R}^{m\times p}$, respectively.

An overparameterized MLR model (3.2) corresponds to an underdetermined linear algebraic system $X\boldsymbol{\beta} = \boldsymbol{y}$, which is usually solved in the minimum-norm least-squares sense as $\hat{\boldsymbol{\beta}} = X^T(XX^T)^{-1}\boldsymbol{y}$, where we have assumed that $\operatorname{rank}(X) = n$. Then, such a 'trained' MLR model will make the following predictions for the testing dataset:

$$\begin{aligned} \hat{\boldsymbol{y}}_\mathrm{t} &= X_\mathrm{t} X^T (XX^T)^{-1} \boldsymbol{y}, \\ \hat{X}_\mathrm{t} &= X_\mathrm{t} X^T (XX^T)^{-1} X. \end{aligned} \tag{3.6}$$

Both $\boldsymbol{y}$ and $X$ are considered to be given in the training dataset $S_n$, whereas only the predictor data-matrix $X_\mathrm{t}$ is considered to be given in the testing dataset $\mathcal{S}_\mathrm{t}$. Hence, strictly speaking, the prediction $\hat{X}_\mathrm{t}$ of the given matrix $X_\mathrm{t}$ is not necessary. However, the fact that it represents a projection of the rows of $X_\mathrm{t}$ on the row space of the training matrix $X$ will be used in our subsequent analysis. Moreover, the two predictions (3.6) can now be written as the prediction $\hat{S}_\mathrm{t}$ of the testing dataset matrix $S_\mathrm{t}$:

$$\hat{S}_\mathrm{t} = X_\mathrm{t} X^T (XX^T)^{-1} S_n. \tag{3.7}$$

The following definition 3.2 establishes the connection between the continuous world of predictor variables as described by the predictor functions and the discrete world of datasets, i.e., the columns of the data matrices.

**Definition 3.2.** A linear vector space $\mathcal{V}_n$, with $\dim(\mathcal{V}_n) = n$, is called a *column function space* of the dataset $\mathcal{S}$ of rank $n$ if there exist $p+1$ *column functions* $x_j(s) \in \mathcal{V}_n$, $x_j : \mathbb{R} \to \mathbb{R}$, $j = 0, \dots p$, such that:

$$\begin{aligned} y_i &= x_0(s_i), \\ x_{i,j} &= x_j(s_i), \quad j = 1, \dots, p, \end{aligned} \tag{3.8}$$

for any data-point $(y_i, x_{i,1}, \dots, x_{i,p})$ from $\mathcal{S}$.

Notice, we use the same notation for the predictor data $x_{i,j}$ and the column functions $x_j(s)$, with the relation being $x_{i,j} = x_j(s_i)$. Also, obviously, $s = y$ in the IR model (3.4).

Let $\{v_k : \mathbb{R} \to \mathbb{R} |\, k = 1, \dots, n\}$ be a basis of the column function space $\mathcal{V}_n$ of the dataset $\mathcal{S}$ of rank $n$. Then, each column function $x_j(s)$ can be expanded as:

$$x_j(s) = \sum_{k=1}^{n} a_{k,j} v_k(s). \tag{3.9}$$

The entries of the *basis matrix* $V \in \mathbb{R}^{n \times n}$ are the sampled basis functions $v_k(s_i)$, $k = 1, \dots, n$, $i = 1, \dots, n$:

$$V = \begin{bmatrix} v_1(s_1) & \dots & v_n(s_1) \\ \vdots & \ddots & \vdots \\ v_1(s_n) & \dots & v_n(s_n) \end{bmatrix}. \tag{3.10}$$

If $V$ is invertible, then the data-vector $\boldsymbol{y}$ and the data-matrix $X$ of the training set $S_n$ can be decomposed as follows:

$$X = VA, \quad \boldsymbol{y} = V\boldsymbol{a}_0, \tag{3.11}$$

where the elements $[A]_{k,j} = a_{k,j}$ of the matrix $A \in \mathbb{R}^{n \times p}$ and the elements $[\boldsymbol{a}_0]_k = a_{k,0}$ of the vector $\boldsymbol{a}_0 \in \mathbb{R}^n$ are the expansion coefficients from (3.9).

Formally, the PARCUR model (3.4) can be trained by computing the matrix $A$ and the vector $\boldsymbol{a}_0$ as:

$$A = V^{-1}X, \quad \boldsymbol{a}_0 = V^{-1}\boldsymbol{y}. \tag{3.12}$$

To predict the test dataset $\mathcal{S}_\mathrm{t}$, one has to solve the following minimization problem:

$$\hat{V}_\mathrm{t} = \arg \min_{V_\mathrm{t} \in \mathbb{R}^{m \times n}} \|X_\mathrm{t} - V_\mathrm{t}A\|_2^2, \tag{3.13}$$

where $X_\mathrm{t}$ is the given predictor test data-matrix. If $\mathrm{rank}(A) = \mathrm{rank}(S_n) = n$, the solution of the problem (3.13) is given by:

$$\hat{V}_\mathrm{t} = X_\mathrm{t}A^T(AA^T)^{-1}. \tag{3.14}$$

Then, applying the relations (3.11), the predictions of the PARCUR model can be written as follows:

$$\begin{aligned}\hat{\boldsymbol{y}}_{\mathrm{t}} &= \hat{V}_{\mathrm{t}}\boldsymbol{a}_0 = X_{\mathrm{t}}A^T(AA^T)^{-1}\boldsymbol{a}_0,\\ \hat{X}_{\mathrm{t}} &= \hat{V}_{\mathrm{t}}A = X_{\mathrm{t}}A^T(AA^T)^{-1}A.\end{aligned} \tag{3.15}$$

This can also be combined into the prediction $\hat{S}_{\mathrm{t}}$ of the testing dataset matrix $S_{\mathrm{t}}$:

$$\hat{S}_{\mathrm{t}} = X_{\mathrm{t}}A^T(AA^T)^{-1}A_n, \tag{3.16}$$

where $A_n = [\boldsymbol{a}_0, A]$ and, from (3.11), $S_n = VA_n$.

The following theorem 3.3 establishes the condition under which the prediction made by the MLR is going to be exact.

**Theorem 3.3.** *The prediction $\hat{S}_{\mathrm{t}}$ produced by the MLR model is exact for any data matrix $S_{\mathrm{t}}$ from the fundamentally overparameterized dataset $\mathcal{S}$ of rank $q$ if and only if the training dataset $S_q$ is complete and* $\mathrm{rank}(X) = q$.

*Proof.* Let $S_q = [\boldsymbol{y}, X]$ be a complete training set of the FOP dataset $\mathcal{S}$. By definition 3.1, for any test set $\mathcal{S}_{\mathrm{t}} \subset \mathcal{S}$ with the data matrix $S_{\mathrm{t}} = [\boldsymbol{y}_{\mathrm{t}}, X_{\mathrm{t}}]$ there exists the matrix $U$ such that $S_{\mathrm{t}} = US_q$ and $X_t = UX$. Hence, from (3.7),

$$\hat{S}_{\mathrm{t}} = X_{\mathrm{t}}X^T(XX^T)^{-1}S_q = UXX^T(XX^T)^{-1}S_q = US_q = S_{\mathrm{t}}. \tag{3.17}$$

If the training set $S_q$ is not complete, then there exists a data row $\boldsymbol{s}_{\mathrm{t}} \in \mathcal{S}$ that is not a linear combination of the rows of $S_q$. □

Whether the training dataset is compete or not, the MLR and the PARCUR models are equivalent in the following sense.

**Theorem 3.4.** *Let $S_n = [\boldsymbol{y}, X]$,* $\mathrm{rank}(X) = n$, *and $\mathcal{S}_{\mathrm{t}} = [\boldsymbol{y}_{\mathrm{t}}, X_{\mathrm{t}}]$ be a training and a testing datasets of the fundamentally overparameterized dataset $\mathcal{S}$ of rank $q$, $n \leq q$. Let also $\mathcal{V}_q$ be the column function space of $\mathcal{S}$. Then, for any basis $\{v_j\}$ in $\mathcal{V}_q$ with the invertible basis matrix $V \in \mathbb{R}^{n\times n}$, the predictions of the testing dataset (3.6) and (3.15) made, respectively, by the MLR and the PARCUR models are equal.*

*Proof.* The proof follows from substituting the relations (3.12) into (3.15). □

If the training dataset is incomplete and the testing data-matrix $S_{\mathrm{t}}$ is not (yet) in the row-range of the training data-matrix $S_n$, there will be an error in the predictions made by the MLR and PARCUR models. Let $S_{\mathrm{t}}$ be decomposed as

$$S_{\mathrm{t}} = US_n + S_{\mathrm{t}}^{\perp}, \quad S_{\mathrm{t}}^{\perp}S_n^T = \boldsymbol{y}_{\mathrm{t}}^{\perp}\boldsymbol{y}^T + X_{\mathrm{t}}^{\perp}X^T = 0_{k,n}, \tag{3.18}$$

where $S_{\mathrm{t}}^{\perp} = [\boldsymbol{y}_{\mathrm{t}}^{\perp}, X_{\mathrm{t}}^{\perp}]$, and $0_{k,n} \in \mathbb{R}^{k\times n}$ is the matrix of all zeros. The existence of this decomposition stems from the fact that $\mathrm{rank}(S_n) = n$, i.e., $S_nS_n^T$ is invertible.

Then, $X_\mathrm{t} = UX + X_\mathrm{t}^\perp$, and the prediction error will be:

$$\begin{aligned} S_\mathrm{t} - \hat{S}_\mathrm{t} &= US_n + S_\mathrm{t}^\perp - (UX + X_\mathrm{t}^\perp)X^T(XX^T)^{-1}S_n \\ &= S_\mathrm{t}^\perp + X_\mathrm{t}^\perp X^T(XX^T)^{-1}S_n = S_\mathrm{t}^\perp - \boldsymbol{y}_\mathrm{t}^\perp \boldsymbol{y}^T(XX^T)^{-1}S_n \\ &= \left[\boldsymbol{y}_\mathrm{t}^\perp - \boldsymbol{y}_\mathrm{t}^\perp \boldsymbol{y}^T(XX^T)^{-1}\boldsymbol{y},\ X_\mathrm{t}^\perp - \boldsymbol{y}_\mathrm{t}^\perp \boldsymbol{y}^T(XX^T)^{-1}X\right]. \end{aligned} \tag{3.19}$$

Hence, apart from the obvious case $\boldsymbol{y}_\mathrm{t}^\perp = \mathbf{0}$, the prediction of the dependent variable will also be exact if $\boldsymbol{y}^T(XX^T)^{-1}\boldsymbol{y} = 1$.

Technically, the difference between the MLR and PARCUR models can be expressed as a simple transformation of the columns of the training dataset $S_n$ by the basis matrix $V$. The relative advantage of the IR version, $s = y$, of the PARCUR model over other choices of the parameter $s$ stems from the direct interpretation of the coefficient matrix $A$ since it provides an insight into the dependence of each predictor variable on the dependent variable. The magnitude and the sign of the coefficients in $A$ reflect the type (e.g., linear, quadratic, etc) and the strength of these dependencies. It is also clear why interpreting the $\boldsymbol{\beta}$ vector of the MLR model might be more problematic, as its relation to the coefficient matrix, $\hat{\boldsymbol{\beta}} = A^T(AA^T)^{-1}\boldsymbol{a}_0$, is rather convoluted. In the standard MLR formulation (3.2), a component $\hat{\beta}_j$ of the vector $\hat{\boldsymbol{\beta}}$ is interpreted as the weight of the additive contribution by the corresponding predictor $x_j$. However, it does not further specify the nature of the mathematical relation between $y$ and $x_j$.

It is also possible, and sometimes profitable, to train the PARCUR model on an *overcomplete* overparameterized training data set $S_n \in \mathbb{R}^{n\times(p+1)}$, where $\mathrm{rank}(S_n) = q \leq p$, but $n > q$. However, since the basis matrix $V \in \mathbb{R}^{n\times q}$, $\mathrm{rank}(V) = q$, is now singular, one has to resort to the Ordinary Least-Squares (OLS) solution of the training problem:

$$\hat{A}_n = (V^TV)^{-1}V^TS_n. \tag{3.20}$$

In that case, $\hat{A} = (V^TV)^{-1}V^TX$, and the prediction of the test set makes use of these OLS estimates as $\hat{S}_\mathrm{t} = X_\mathrm{t}\hat{A}^T(\hat{A}\hat{A}^T)^{-1}\hat{A}_n$. Although, the equivalence to the MLR model is only achieved if the chosen basis matrix $V$ coincides with the matrix of the left singular vectors of $X$.

## 3.3 Data containing polynomial column functions

theorem 3.3 shows that the prediction by the MLR model is exact if the training set is complete. From the definition 3.1 it is clear that a complete dataset is a subset of a FOP dataset, such that its data matrix has the maximal possible rank $q$, which can be smaller than the number of predictors $p$. Given a finite training dataset of size $n$ it is hard to tell if: a) it is a subset of a FOP dataset with some $q < p$, b) it is a complete dataset, i.e. $n = q$. In this section, working under the assumption that the data set $\mathcal{S}_m$ does not contain statistical noise, we establish the sufficient condition for the existence of a FOP dataset. To formulate these conditions it is convenient to make a choice of the column function space, see definition 3.2. Here, we consider the polynomial function space and the *monomial* basis, which lead to easily interpretable regression results.

We note a well-known fact that the monomial basis $\{v_k(y) = y^k \,|\, k = 0, \ldots, n-1\}$ for the column function space $\mathcal{V}_n$ will produce an invertible Vandermonde basis matrix $V$, given by

$$V = \begin{bmatrix} 1 & y_1 & y_1^2 & \cdots & y_1^{n-1} \\ 1 & y_2 & y_2^2 & \cdots & y_2^{n-1} \\ \vdots & \vdots & \cdots & \vdots & \\ 1 & y_n & y_n^2 & \cdots & y_n^{n-1} \end{bmatrix}, \tag{3.21}$$

if all entries of the dependent variable data-vector $\boldsymbol{y} = [y_1, \ldots, y_n]^T$ are distinct. In this basis, the entries $a_{i,j}$ of the coefficient matrix $A$ represent the coefficients of the polynomial functions $x_j(y)$ that may generate some of the columns of the predictor data-matrix $X$:

$$x_j(y) = a_{1,j} + a_{2,j}y + \cdots + a_{n,j}y^{n-1}. \tag{3.22}$$

This also puts the PARCUR model into the context of polynomial fitting. Whether the columns of $X$ have or have not been generated by polynomial functions of $y$, the IR model with the basis matrix (3.21) will be projecting all columns of $X$ on the monomial basis. From the computational point of view, Vandermonde matrices, while theoretically invertible, are hard to work with for sizes above $n = 15$ and become the source of significant round-off errors. Luckily, one normally does not need polynomial functions of very high degree to adequately describe a data column. To minimize the numerical errors, we also normalize the range of the $y$-data to fit within the interval $[-1, 1]$.

In general, it is difficult to decide whether the training dataset is complete by simply inspecting the entries of its data matrix $S_n$. In theory, one could compute the Singular-Value Decomposition (SVD) of the matrix and see if the zero singular value appears after adding any new sample (row) to the training dataset. However, here we are interested in an arbitrary column function space $\mathcal{V}_q$ with the square invertible basis matrix $V \in \mathbb{R}^{n\times n}$, and the conclusions about the eventual completeness of $S_n$ will be based on the shape of the corresponding coefficient matrix $A_n = V^{-1}S_n$.

Each of the $p+1$ data columns $x_j$ in a dataset $\mathcal{S}$ belongs to one of the three classes:

1. $x_j(y)$ is a polynomial in $y$ of degree less or equal to $n-1$
2. $x_j(y)$ is a polynomial in $y$ of degree higher than $n-1$
3. there is a non-functional dependence between $x_j$ and $y$

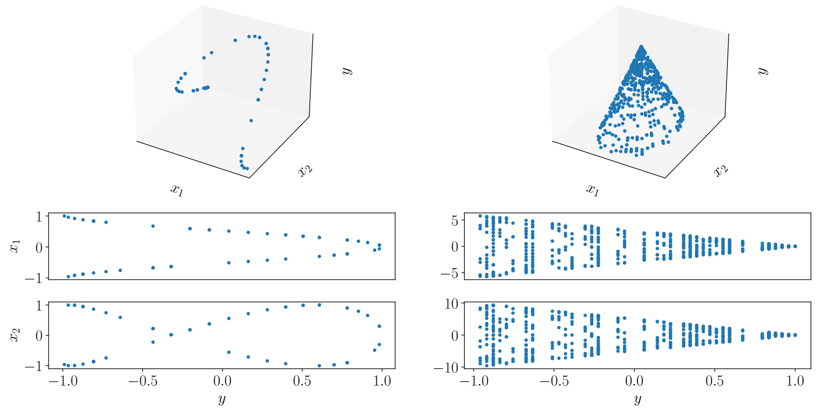


Figure 3.2: Examples of column data produced by non-functional relationships between the dependent variable $y$ and the predictor variables $x_1$ and $x_2$. Left: data on a curve that cannot be parameterized by $y$. Right: data on a conical surface. Two-dimensional scatter-plots: $x_1$ and $x_2$ column data sorted by $y$ and displayed as 'functions' of $y$.

In particular, the first column $x_0 = y$, i.e., the dependent variable, obviously, belongs to the first class with $n = 2$.

A non-functional dependence between the predictor $x_j$ and the dependent variable $y$ would emerge if the data was situated on a curve that cannot be parameterized by $y$, fig. 3.2 (left). Choosing a different parameterization could transform these 'nonfunctional' data to functions $y(s)$, $x_j(s)$, $j = 1, \ldots, p$, as in the general PARCUR formulation. A more severe case of non-functional dependence arises where the data is situated on a higher-dimensional manifold, such as a hyper-surface, fig. 3.2 (right), and no alternative parameterization can fix this problem. The column data that one observes in such 'nonfunctional' cases are illustrated in fig. 3.2 (bottom, two-dimensional scatter plots).

**Theorem 3.5.** *Let $\mathcal{S}$ be a dataset with one independent variable $y$ and $p$ predictor variables $x_j$, $j = 1, \ldots, p$. Let also $\mathcal{S}$ contain $k \leq p$ predictors that are polynomials in $y$ of degrees greater than $q - 1$, or have a non-functional relation to $y$. Then, $\mathcal{S}$ is a fundamentally overparameterized dataset (in linear sense) of rank $q$, $2 \leq q \leq p$, if its remaining $p - k$ predictors are the polynomials in $y$ of degree $r$, $1 \leq r \leq q - k - 1$.*

*Proof.* Without the loss of generality we may assume that for any subset of size $m$ of the dataset $\mathcal{S}$, the dependent variable data $\boldsymbol{y}$ contains only the distinct values of $y$ so that the square monomial basis matrix $V \in \mathbb{R}^{m \times m}$ is invertible. Then, any data-matrix $S_m = [\boldsymbol{y}, X] \in \mathbb{R}^{m\times(p+1)}$ can be represented as $S_m = VA_m$. By the conditions of the Theorem, for any $m > q$, subject to column reordering, the coefficient matrix $A_m \in \mathbb{R}^{m\times(p+1)}$ has the structure:

$$A_m = \begin{bmatrix} \boldsymbol{e}_2 & A_{1,1} & A_{1,2} \\ \boldsymbol{0} & O & A_{2,2} \end{bmatrix}, \; A_{1,1} \in \mathbb{R}^{(q-k)\times(p-k)}, \; A_{1,2} \in \mathbb{R}^{(q-k)\times k}, \; A_{2,2} \in \mathbb{R}^{(m-q+k)\times k}, \tag{3.23}$$

where $V^{-1}\boldsymbol{y} = \boldsymbol{e}_2 \in \mathbb{R}^m$ is the second standard basis vector and $O \in \mathbb{R}^{(m-q+k)\times(p-k)}$ is the matrix of all zeros. Here we have used the fact that the polynomial fit to non-functional data may produce a polynomial of degree greater than $q-1$. It is obvious that $\mathrm{rank}(A_{1,1}) \leq (q-k)$ and, for any $m \geq (q-k)$, $\mathrm{rank}(A_{2,2}) \leq k$. Therefore, $\mathrm{rank}(A) \leq q$, and $\mathrm{rank}(S_m) \leq q$ for any $m$, showing that $\mathcal{S}$ is a FOP dataset. □

The above theorem 3.5 shows that having extremely high-degree polynomials, e.g., with degrees higher than $p-1$ and non-functional dependencies among the predictors does not prevent the MLR and the IR models from making exact predictions as long as there are also polynomial data of sufficiently low degree and the set on which the model is trained is complete. Moreover, a complete dataset can, in principle, be achieved with $n < p$ samples. The latter fact may seem surprising as it appears that we are able to recover a polynomial of degree higher than $n-1$ or a non-functional dependence by training on just $n$ data points. However, it becomes less surprising if we consider the form of the MLR and IR predictors given by (3.6) and (3.15), as in both cases the leftmost matrix $X_{\mathrm{t}}$ contains the $X$-data from the test dataset which ones is trying to 'predict'.

In the limiting case with $p$ high-degree polynomials and/or non-functional dependencies, a complete dataset will only be achieved with $n = p$ samples. It seems to be a waste of time and resources, though, to collect so much training data knowing that the majority of predictors do not even satisfy the model assumptions. We come back to this question in Section 3.5. In the other limiting case, where all predictors are polynomials, the size of the complete training dataset can be as small as $n = 2$ if the maximal degree of all polynomials is at most one and there is at least one predictor which is a linear function of $y$.

Figure 3.3 (top) illustrates the performance of the IR model with the three classes of column data discussed above. Specifically, we are considering the cases where: all predictors are polynomial functions in $y$ of degree $r-1 \leq q$ (blue, circles), some of the predictors are high-degree polynomials in $y$ (orange, squares), and some of the predictors have non-functional relation to the dependent variable $y$ (green, triangles).

In these numerical experiments the data $x_j(y)$ are generated by randomly sampling the range of $y \in [-1, 1]$ and evaluating polynomial functions of various degrees at the sampled points. Columns that are not functions of $y$ are generated as non-invertible functions $y(x_j)$, similar to those in the examples of fig. 3.2. The predictions are

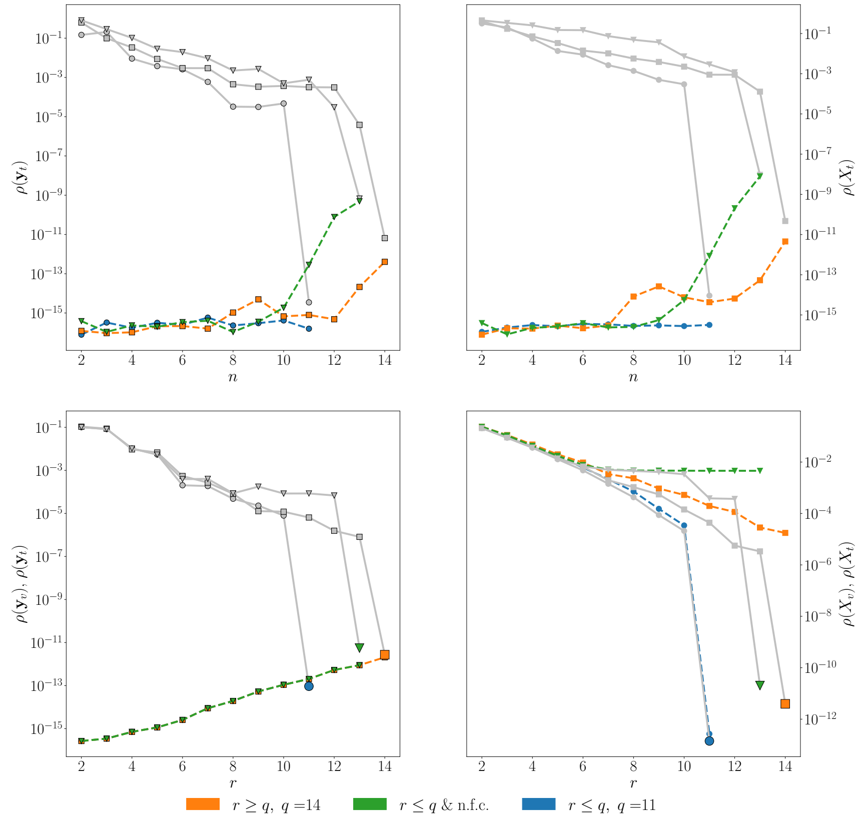


Figure 3.3: Top: prediction errors of the IR model as functions of the training dataset rank obtained with invertible basis matrix $V$ and exact data (see text for full explanation). Bottom: test of the regularization algorithm on exact (noiseless) data and over-complete training dataset. Bright enlarged colored markers correspond to the minima of the validation errors that indicate the optimal polynomial degree $r^*$ along the horizontal axis (bottom, right) and show the values of the test errors attained with these $r^*$ (bottom, left). Note, the horizontal axis displays $\mathrm{rank}(\hat{A}) = r + 1$, rather than the actual polynomial degree $r$.

computed as in Eq.'s (3.15), where the coefficient matrix $A$ is obtained from the training $X$-data as in Eq. (3.12).

The ranks of the complete datasets are: $q = 11$ ($p = 200$ polynomials of degree $r \leq 10$), $q = 14$ ($p = 203$, with 200 polynomials of degree $r \leq 10$ and three polynomials of degree $r = 17$), and $q = 13$ ($p = 202$, with 200 polynomials of degree $r \leq 10$ and two non-functional predictors). In all three cases, we expect the prediction errors to vanish as soon as the rank of the training dataset reaches $q$.

The prediction errors in $\boldsymbol{y}$ and $X$ are measured on both the training and the test sets as follows:

$$\begin{aligned} \rho(\boldsymbol{y}_{\mathrm{v}}) &= \frac{\|\boldsymbol{y}_{\mathrm{v}} - \hat{\boldsymbol{y}}_{\mathrm{v}}\|_2}{\|\boldsymbol{y}_{\mathrm{v}}\|_2}, \quad \rho(X_{\mathrm{v}}) = \frac{\|X_{\mathrm{v}} - \hat{X}_{\mathrm{v}}\|_F}{\|X_{\mathrm{v}}\|_F}, \\ \rho(\boldsymbol{y}_{\mathrm{t}}) &= \frac{\|\boldsymbol{y}_{\mathrm{t}} - \hat{\boldsymbol{y}}_{\mathrm{t}}\|_2}{\|\boldsymbol{y}_{\mathrm{t}}\|_2}, \quad \rho(X_{\mathrm{t}}) = \frac{\|X_{\mathrm{t}} - \hat{X}_{\mathrm{t}}\|_F}{\|X_{\mathrm{t}}\|_F}, \end{aligned} \tag{3.24}$$

where $\boldsymbol{y}_{\mathrm{v}}$ and $X_{\mathrm{v}}$ are the training (validation) set data and $\boldsymbol{y}_{\mathrm{t}}$ and $X_{\mathrm{t}}$ are the test set data.

fig. 3.3 shows the prediction errors as functions of the training set rank (top row), and of the polynomial representation degree (bottom row). The bright colored dashed lines give the training set errors and the dim solid gray lines show the corresponding errors on the test dataset. Plots on the left show the dependent variable errors and on the right – the errors for the predictor variables. As expected, the errors drop as soon as the training dataset becomes complete. With the high-degree polynomial predictors and especially with non-functional predictors the errors on the training set begin to rise at the end and the errors on the test set do not drop to machine precision values as happens with the low-degree polynomial predictors. This is due to the fact that the coefficient matrices in the former two cases become ill-conditioned, making the predictors more sensitive to numerical round-off errors.

## 3.4 Polynomial regularization

All measured data are either random in nature or contain statistical noise. This is already implicitly assumed in eq. (3.1) which interprets the observed $y$ as a realization of an unknown random variable $Y$. Indeed one usually measures the quantity $y_i + \epsilon_i$, where the 'noise' $\epsilon_i$ is a single realization of a random variable $\epsilon$ with the distribution from some well-defined class, e.g., $\epsilon \sim \mathcal{N}(0, \sigma^2)$, see e.g. [37]. The role of the regression model is to recover the parameters of the distribution generating $y$, such as the mean and the variance $\sigma^2$, which are unknown and estimated from the data. Being measured quantities, the independent variables $x_j$ may also be contaminated by noise. In the noisy setting, the PARCUR model becomes:

$$\begin{aligned} y &= y(s) + \epsilon_0, \\ x_j &= x_j(s) + \epsilon_j, \quad j = 1, \dots, p; \\ s &\in [a, b] \subset \mathbb{R}. \end{aligned} \tag{3.25}$$

For simplicity we assume here the most common statistical hypotheses $\epsilon_j \sim \mathcal{N}(0, \sigma_j^2)$, $j = 0, \dots, p$, which provides an adequate description of the "standard" errors due to the finite sample size. We investigate the effect of the additive Gaussian noise $\epsilon_j \sim \mathcal{N}(0, \sigma_j^2)$, $j = 0, \dots, p$ applied in the following three ways:

1. noise only in the dependent variable $y$: $\sigma_0 = \sigma \neq 0$, $\sigma_j = 0$, $j = 1, \dots, p$
2. independent and identically distributed (i.i.d.) noise in $S_n$: $\sigma_j = \sigma \neq 0$, $j = 1, \dots, p$
3. independent, but not identically distributed noise in $S_n$:

$$\sigma_j = \begin{cases} \sigma \neq 0, & \text{if } j = 0 \text{ and } x_j \in \{\text{noisy predictors}\} \\ 0, & \text{if } j = 0 \text{ and } x_j \in \{\text{exact predictors}\} \end{cases} \tag{3.26}$$

Obviously, the first 'classical' case belongs to the last class if considered over the complete data matrix $S_n$.

To avoid overfitting and mitigate the effect of noise on the model predictions, one can use any of the standard regularization techniques, such as the Tikhonov regularization or the truncated SVD. In the case of the IR model, truncating the number of terms of the monomial basis used to represent the column functions appears to be the most natural regularization approach.

In the previous section 3.2, the exact training data $\boldsymbol{y}$, could simply be sorted by magnitude and used as the values of the curve parameter to construct the monomial basis matrix $V$. While this is still permitted in the noisy case (we can choose the parameter $s$ as we wish), the interpretation of the coefficient matrix $A$ is no longer straightforward, if instead of $\boldsymbol{y}$ we are using $\boldsymbol{y} + \boldsymbol{\epsilon}$.

At the risk of losing some of the interpretability of the coefficient matrix $\hat{A}$, we shall nevertheless be sorting all our data by the magnitude of $\boldsymbol{y} + \boldsymbol{\epsilon}$. Notice, that the vector $\boldsymbol{y} + \boldsymbol{\epsilon}$ sorted by the (inaccessible) exact data $\boldsymbol{y}$ represents a smooth function (irregularly sampled $y$) with an additive 'white' noise. Sometimes, sorting by the exact $\boldsymbol{y}$ can be achieved if the predictors contain a 'clean' (noiseless) monotonous function of $y$ or even a noisy monotonous function that has been sorted in correct order. In that case, all columns in $S_n$ should be sorted by the column corresponding to this predictor variable.

If $\boldsymbol{y} + \boldsymbol{\epsilon}$ is correctly sorted by $\boldsymbol{y}$, then a multitude of denoising techniques is available for such smooth noisy signals, e.g., the Wiener filter. However, numerical experiments indicate that, in our case, the Wiener filter becomes effective starting from approximately $n = 150$ data samples and is of little use below that threshold. Similar lower bound seems to hold for the effectiveness of the cross-validation and similar techniques that allow to deduce the optimal value of the regularization parameter (maximal degree $r^*$ of the monomial basis functions) from the training data. Therefore, we propose a regularization approach that depends on the number of available training samples, see table 3.1.

The optimally regularized representation $\hat{\boldsymbol{x}}_j(r^*)$ of a noisy data-column $\boldsymbol{x}_j + \boldsymbol{\epsilon}_j$ minimizes the error between the exact (noiseless) column $\boldsymbol{x}_j$ and its regularized representation $\hat{\boldsymbol{x}}_j(r^*)$, e.g. $\|\boldsymbol{x}_j - \hat{\boldsymbol{x}}_j(r^*)\|_2$, where $r^*$ is the truncation index in terms of the monomial basis.

| | $n < 150$ | $n \geq 150$ |
|---|---|---|
| Step 1 | Sort $\boldsymbol{y} + \boldsymbol{\epsilon}$ (if possible by $\boldsymbol{y}$) | |
| Step 2 | – | Apply Wiener filter on sorted $\boldsymbol{y} + \boldsymbol{\epsilon}$ |
| Step 3 | Sort all columns by sorted $\boldsymbol{y} + \boldsymbol{\epsilon}$ | Sort columns by sorted and filtered $\boldsymbol{y} + \boldsymbol{\epsilon}$ |
| Step 4 | Choose $r^*$ subjectively | Use CV to find $r^*$ |

Table 3.1: Regularization strategies for the IR model with truncated monomial basis, depending on the number of training samples.

Since the noiseless column is not known, the optimal truncation index $r^*$ is determined either subjectively by observing the quality of fit for several columns ($n < 150$) or by the Cross Validation (CV) technique ($n \geq 150$).

We employ a 10-fold CV, where during each fold the model is trained on a randomly chosen subset of the training dataset and evaluated on a complementary (validation) subset. As a metric, we consider the following validation errors:

$$\rho(\boldsymbol{y}_{\mathrm{v}}) = \left\langle \frac{\|\boldsymbol{y}_{\mathrm{v}} - \hat{\boldsymbol{y}}_{\mathrm{v}}(r)\|_2}{\|\boldsymbol{y}_{\mathrm{v}}\|_2} \right\rangle, \qquad \rho(X_{\mathrm{v}}) = \left\langle \frac{\|X_{\mathrm{v}} - \hat{X}_{\mathrm{v}}(r)\|_{\mathrm{F}}}{\|X_{\mathrm{v}}\|_{\mathrm{F}}} \right\rangle, \tag{3.27}$$

where the angular brackets denote the arithmetic averaging over the CV folds. The optimal truncation index $r^*$ corresponds to the polynomial degree for which the minimum of $\rho(X_{\mathrm{v}})$ is attained.

There is a curious reason behind the fact that we have to use the error $\rho(X_{\mathrm{v}})$ in the $X$-data rather than the usual error $\rho(\boldsymbol{y}_{\mathrm{v}})$ in the $\boldsymbol{y}$-data. In the polynomial IR method, the vector of the dependent variable (either exact or noisy) is the second column of the basis Vandermonde matrix $V$, see (3.21). Therefore, in exact arithmetic, the training $\boldsymbol{y}$-data is exactly reproduced for any $r \geq 2$. Hence, strictly speaking, the IR model always over-fits the training $\boldsymbol{y}$-data. The regularization of the IR model is validated and tuned on the columns of the matrix $X$. This is possible, since apart from the noise in $X$ itself, the $\boldsymbol{y}$-noise is always propagated into the Vandermonde matrix $V$ and then into the columns of the coefficient matrix $\hat{A} = (V^T V)^{-1} V^T X$. Thus, the prediction $\hat{X}_{\mathrm{t}} = X_{\mathrm{t}} \hat{A}^T (\hat{A}\hat{A}^T)^{-1} \hat{A} X$ will be always affected by noise. Simply put, using the noisy data $\boldsymbol{y} + \boldsymbol{\epsilon}$ to construct $V$ is equivalent to using a wrong parameterization $s \neq y$, whereas the column data are generated in the $s = y$ parameterization.

Finally, it is important to realize that there are two ways to regularize the noisy data. One, which is the focus of the present section, is to choose a single optimal (maximal) degree $r^*$ for the representation of all data columns, i.e., both $\boldsymbol{y}$ and all $\boldsymbol{x}_j$, $j = 1, \dots, p$. This, obviously, has its drawbacks, since some of the columns may be noiseless or contain a different level of noise. A more flexible and precise way is to determine the optimal degree $r_j^*$, $j = 0, 1, \dots, p$, for each column individually. Our preliminary numerical experiments have shown that the latter ‘flexible’ regularization approach does not necessarily result in a better prediction of the test $\boldsymbol{y}_{\mathrm{t}}$ data.

Nonetheless, in our opinion, this flexible regularization deserves a separate in-depth investigation in the case of the non-i.i.d noise and for the purposes described in the next section 3.5.

fig. 3.3 (bottom) illustrates the testing of the polynomial-degree truncation regularization scheme for the IR model on the three noiseless datasets described in section 3.3. The meaning of the lines and symbols is the same as in fig. 3.3 (top), see section 3.3. The errors $\rho(\boldsymbol{y}_\mathrm{v})$ (left, colored, dashed) and $\rho(X_\mathrm{v})$ (right, colored, dashed) on the training dataset are the mean CV errors (3.27). All data are exact up to numerical precision. The only differences with the IR model of fig. 3.3 (top) are the overcomplete nature of the training datasets ($n = 150$) and the application of the LS estimates (3.20) of the coefficient matrix $\hat{A}_r$. Therefore, the horizontal axis in fig. 3.3 is the number $r$ of columns in the rectangular basis matrix $V \in \mathbb{R}^{n\times r}$, which is no longer equal to the number of samples $n$.

In the tests of fig. 3.3 (bottom-left), the $\rho(\boldsymbol{y}_\mathrm{v})$ error starts at the level of machine precision for $r = 2$ and increases thereafter due to the accumulation of numerical errors. That the $\rho(\boldsymbol{y}_\mathrm{v})$ error grows for $r > 2$ has to do with the fact that the exact representation of $\boldsymbol{y}$ is attained already at $r = 2$ in the monomial basis. The enlarged coloured markers in the right plot indicate the minima of the $\rho(X_\mathrm{v})$, the same markers in the left plot show the levels of $\rho(\boldsymbol{y}_\mathrm{t})$ errors attained with the corresponding $r^*$. As can be seen, the optimal $r^*$ gives the smallest achievable error $\rho(\boldsymbol{y}_t)$. Hence, in this noiseless case the regularization procedure correctly identifies the rank of the complete dataset as the optimal $r^*$.

In the next numerical experiments, for training, validation and testing, we use a synthetic dataset with $p = 202$ predictors. Two of these 202 predictors are generated by randomly sampling a non-functional relationship such as those shown in Figure 3.2. One of these 'non-functional' columns has the data from a curve that cannot be parameterized by $y$, another has the data from a cone. The remaining 200 features are obtained by evaluating polynomials of degree lower than 12 at the sample points $y_i$. While we have investigated all three types of noise listed above Eq. (3.26), we only present the results for the first and the third cases, as the second, i.i.d. case appeared to be very similar to the third, non-i.i.d case. In all examples we use $n = 150$ and the optimal polynomial degree $r^*$ is found with the CV method at the minimum of $\rho(X_\mathrm{v})$.

fig. 3.4 (top) illustrates the application of the regularization procedure to the problem where only the dependent-variable data $\boldsymbol{y}$ contains additive Gaussian noise. The columns of the matrix $X$ are exact up to machine precision. We consider three different standard deviations $\sigma = 0.05, 0.1$, and $0.2$, corresponding to 5%, 10%, and 20% levels of noise relative to the $\boldsymbol{y}$-data magnitude. As can be seen from the test-error curves (top-left plot, dim solid gray), the errors $\rho(\boldsymbol{y}_\mathrm{t})$ attained with these choices of $r^*$ (top-right plot, enlarged colored squares) are close to the smallest achievable $\rho(\boldsymbol{y}_\mathrm{t})$ errors for all considered noise levels. At the same time, the minimal $\rho(X_\mathrm{v})$ errors do not correspond to the smallest achievable $\rho(X_\mathrm{t})$ errors (top-right plot, dim solid gray curves).

fig. 3.4 (bottom) illustrates the case of additive noise in the dependent variable and in one-third of the predictors. While the behavior is generally similar to the previous case, the attained optimal $\rho(\boldsymbol{y}_{\mathrm{t}})$ errors increase more rapidly with the level of noise. The growth of the $\rho(\boldsymbol{y}_{\mathrm{v}})$ error (bottom-left, colored dashed) is now caused not only by the accumulation of numerical errors and the statistical noise in $\boldsymbol{y}$ (via $V$), but by the statistical noise in the $X$-data as well (via $\hat{A}$).

Finally, we remark that the error $\rho(\boldsymbol{y}_{\mathrm{v}})$ attained with the optimal $r^*$ on the training-validation dataset is always much smaller (see the range of the right vertical axis) than the corresponding error $\rho(\boldsymbol{y}_{\mathrm{t}})$ on the test dataset. This is, obviously, one more case of the benign overfitting [11], [32], caused by the choice of the monomial column basis in or IR model rather than the statistical properties of the noise.

## 3.5 Removal of improper predictors

From the analysis and examples of the previous sections it is clear that a linear model, be it the classical MLR implementation or the present IR formulation, will produce reasonably exact predictions even if some of the predictor variables do not satisfy the model assumptions. In the regularized IR formulation of section 3.4 the model assumption is that the predictor is a polynomial in the dependent variable $y$ of degree at most $r^* - 1$, where $r^*$ is the rank of the optimally regularized $\hat{A}$. In the case of exact data discussed in section 3.3, the training set may become complete way before some of the predictor functions have been properly sampled. Success of predictions in the presence of such 'improper' predictors gives an illusion of understanding of the underlying natural phenomena [46]. In section 3.4 we have also investigated the case where the noise is present only in some of the predictors. Such noisy predictors may be negatively affecting the predictive power of the model. Thus, there appear to be two good reasons to detect and possibly discard both the high-degree/non-functional and the noisy predictors from the model. Additionally, in biological and agricultural applications, the large number of predictors included in the FOP dataset is often due to a broad (untargeted) experimental search in the absence of any *a-priori* information. The practical goal of such studies is to identify a preferably small subset of microbiota, fungi, molecules, metabolites, or genes, allowing for future targeted and less expensive measurements of these predictor variables.

Discarding 'unnecessary' predictors is known as feature or variable selection in statistics and ML. The main idea of feature selection is simple: one can sometimes achieve the same or even better prediction with just a subset of predictors. However, no clear principle for the inclusion or removal of any particular predictor has been put forward so far. Therefore, feature selection methods are either combinatorial or heuristic in the ways they produce the candidate subsets of predictors. Moreover, the criterion for choosing a particular subset is the prediction error on the training dataset. Since this

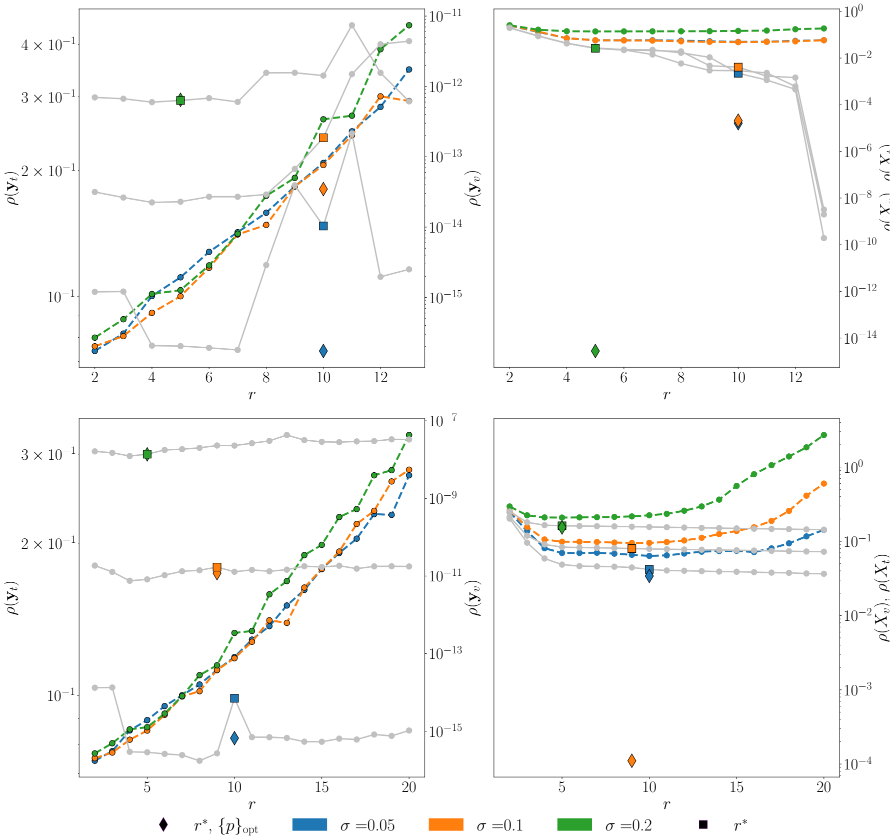


Figure 3.4: Top: application of the regularization procedure in the case of exact predictor data $X$ and noisy dependent-variable data $\boldsymbol{y}+\boldsymbol{\epsilon}$, $\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\sigma_i^2 I)$, with $\sigma_1=0.05$ (blue), $\sigma_2=0.1$ (orange), and $\sigma_3=0.2$ (green). Bright colored lines: validation errors $\rho(\boldsymbol{y}_{\text{v}})$ (left, right vertical axis) and $\rho(X_{\text{v}})$ (right). Dim dashed gray lines: errors on the test dataset, $\rho(\boldsymbol{y}_{\text{t}})$ (left, left vertical axis) and $\rho(X_{\text{t}})$ (right). Bright colored square markers: values of the test errors $\rho(\boldsymbol{y}_{\text{t}})$ and $\rho(X_{\text{t}})$ attained at $r^*$ (left and right respectively). Bright colored diamond markers: values of the test errors $\rho(\boldsymbol{y}_{\text{t}})$ (left, left vertical axis) and $\rho(X_{\text{t}})$ (right) attained with $r^*$ and the optimal set $\{p\}_{\text{opt}}$ of predictors. Bottom: same as in the row above, but with a third of the columns of $X$ affected by the additive Gaussian noise of the same type as the noise in the dependent variable $\boldsymbol{y}$.

error has already been used to find the optimal regularization parameter, the feature selection methods are often regarded with suspicion in statistics [37], Chapter 5.

So far we have tuned one hyper-parameter, the optimal degree $r^*$, to define the

| **hyper-parameter** | **data used** |
|---|---|
| optimal degree $r^*$ | $\rho(X_\mathrm{v})$, training $X$-data |
| which predictor to remove | $\chi_j$, test $X_\mathrm{t}$-data |
| optimal threshold $\tau_\mathrm{opt}$ | $\rho(\boldsymbol{y})$, training $\boldsymbol{y}$-data |

Table 3.2: Hyper-parameters of the regularized polynomial IR model with predictor removal and the datasets used to tune these hyper-parameters.

regularized polynomial IR model. We have used the CV procedure on the training $X$-data to determine this hyper-parameter. This leaves the training $\boldsymbol{y}$-data and the test $X_\mathrm{t}$-data available for tuning other hyper-parameters in our model.

Since the goal is to remove the eventual 'improper' predictors, it is logical to use the prediction of the $X_\mathrm{t}$-data as the 'usefulness' criterion for each predictor. We introduce the column-wise test set prediction error:

$$\chi_j = \frac{\|\hat{\boldsymbol{x}}_j - \boldsymbol{x}_j\|_2}{\|\boldsymbol{x}_j\|_2}, \quad j = 1, \ldots, p, \tag{3.28}$$

where $\boldsymbol{x}_j$ and $\hat{\boldsymbol{x}}_j$ are, respectively, the $j$th columns of the test matrix $X_\mathrm{t}$ and its predictor $\hat{X}_\mathrm{t}$. One would, naturally, like to remove the predictors with large $\chi_j$'s. For example, all predictors with $\chi_j \geq \tau$.

This introduces another hyper-parameter, the threshold $\tau$, which can be tuned utilizing the last available portion of our data – the training $\boldsymbol{y}$-data. Recall that this data could not be used to tune the regularization parameter $r^*$, since $\rho(\boldsymbol{y})$ always grows with $r$. Yet, for a fixed $r = r^*$ the error $\rho(\boldsymbol{y})$ depends on the set of included predictors, and we call it the feature removal error. table 3.2 summarizes the data usage by the regularized polynomial IR algorithm with predictor removal for tuning its hyper-parameters.

We define the optimal threshold $\tau_\mathrm{opt}$ to be the one for which the feature-removal error $\rho(\boldsymbol{y})$ with $r = r^*$ is minimal. Unfortunately, while $\rho(\boldsymbol{y})$ normally tends to grow for small $\tau$, since one starts to loose the useful predictors, the overall dependence of $\rho(\boldsymbol{y})$ on $\tau$ is neither unimodal nor smooth. Meaning, that there may be a set or a range of $\tau_\mathrm{opt}$ that gives approximately the same $\rho(\boldsymbol{y})$ at $r = r^*$. In such cases we suggest to make a subjective choice of $r^*$ favouring the smallest number of retained predictors.

We have applied the predictor removal procedure to the case of 200 polynomial columns with degrees $q - 1 \leq 12$ and two non-functional columns in the data-matrix $X$, where the dependent variable and a third of the columns of $X$ are affected by additive Gaussian noise with $\sigma = 0.05$.

fig. 3.5 presents two examples of the feature removal error on the training $\boldsymbol{y}$-data (solid blue) together with the corresponding non-accessible feature removal errors on the test $\boldsymbol{y}_\mathrm{t}$-data (dashed gray). In the left plot, where the level of noise is $\sigma = 0.05$, the choice of the optimal threshold $\tau_\mathrm{opt}$ from the minimum of the training $\boldsymbol{y}$-data corresponds to the smallest possible number of predictors. In the right plot, where

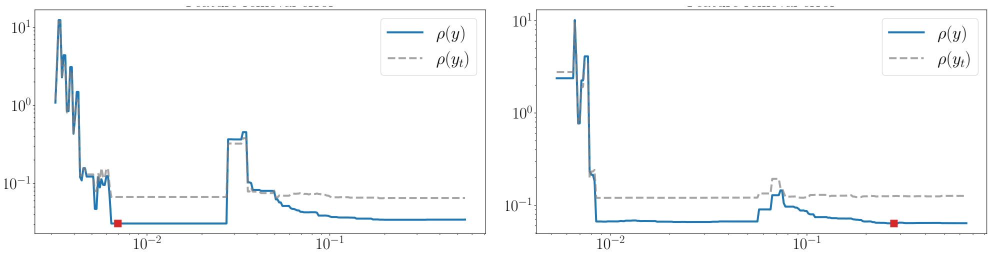


Figure 3.5: Feature removal errors $\rho(\boldsymbol{y})$ (solid blue) and $\rho(\boldsymbol{y}_\mathrm{t})$ (dashed gray), both with $r = r^*$, for synthetic datasets containing additive noise in the dependent variable and one-third of predictors, $\sigma = 0.05$ (left) and $\sigma = 0.1$ (right).On the $y$-axis is the value of the residuals, on the $x$-axis is the value of the threshold $\tau$. Red squares indicate the minima, corresponding to the optimal thresholds $\tau_\mathrm{opt}$.

$\sigma = 0.1$, the detected minimum of the feature removal error is situated in the second valley of the error function and corresponds to many unnecessary predictors retained in the model, as the depth of both valleys is approximately the same. Checking for other local minima in this case and choosing the leftmost along the $\tau$-axis would be a better strategy to find $\tau_\mathrm{opt}$. Regardless, it is comforting to observe that the feature removal errors on the training and test datasets appear synchronous in these numerical examples.

In fig. 3.6 (top) the green triangles show the true ‘degrees’ (lengths of the nonzero parts of the corresponding columns of the coefficient matrix $A$, used to generate the predictor data) for each predictor. The vertical dashed dark-grey lines indicate the predictors containing the Gaussian noise. The vertical dashed red lines indicate the non-functional predictors. The horizontal solid orange line depicts the optimal degree $r^*$.

In fig. 3.6 (bottom) the orange circles show the value of the column prediction error $\rho(\boldsymbol{x}_j)$ on the test $X_\mathrm{t}$-data for each predictor obtained with $r = r^*$. The dashed blue line is the optimal threshold $\tau_\mathrm{opt}$. All predictors with the errors above that line (solid light-gray vertical lines) are discarded. Notice that all noisy predictors, as well as both non-functional predictors are successfully removed.

Diamond markers in fig. 3.4 show by how much the removal of ‘improper’ predictors improves the prediction error $\rho(\boldsymbol{y}_t)$. In fact, it is close to the minimal achievable error with all predictors retained. That level of error was, however, not always obtained with the crude global regularization procedure. Hence, the removal of predictors may sometimes compensate for the errors in determining the optimal regularization parameter.

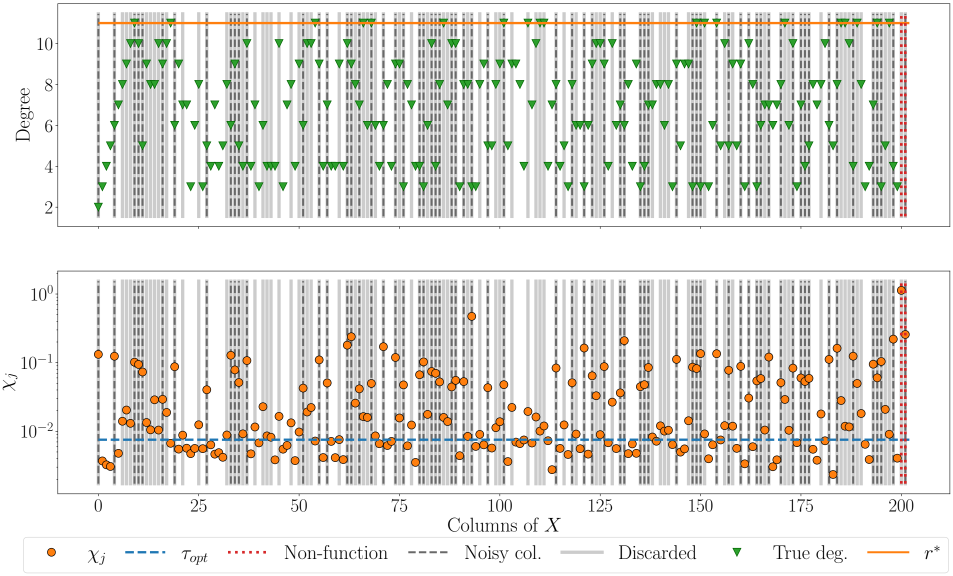


Figure 3.6: Removal of 'improper' predictors (vertical solid light-gray lines) from synthetic data. Top: actual degrees of each predictor (green triangles), the global truncation degree $r^*$ (orange horizontal line), predictors containing Gaussian noise (vertical dashed dark-gray lines), non-functional predictors (vertical dotted red lines). Bottom: $\chi_j$ errors (orange circles), the optimal threshold $\tau_{\text{opt}}$ (horizontal dashed blue line).

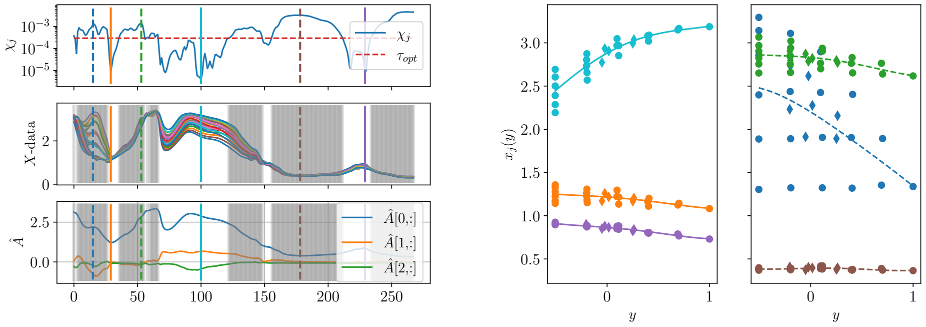
## 3.6 Application to chemometric data

In this section we present the application of the regularized IR model with predictor removal to the experimental dataset that is often used to compare various regularized MLR models, such the Principal Component Regression (PCR) or the Partial Least Squares (PLS) [37]. This dataset was first published in [60] and is also included in the *R*-library `pls` [47] under the name of `yarn`. The dataset consists of 28 Near InfraRed (NIR) spectra of Polyethylene terephthalate (PET) yarns, measured at $p = 268$ wavelengths, and 28 corresponding yarn densities. Seven of these 28 samples are designated as the test dataset and the remaining $n = 21$ constitute the training dataset. The predictor variables (spectral amplitudes) change continuously with the wavelength, which becomes clear if one plots the rows of $X$ as curves, see the middle plot of fig. 3.7a.

The training dataset columns have been sorted by the $\boldsymbol{y}$-data. Since $n = 21 < 150$, we use the visual inspection of the fitted polynomials, see fig. 3.7b, rather than the CV procedure to determine the optimal degree $r^* = 5$. The corresponding validation error $\rho(X_{\mathrm{v}})$ is shown in fig. 3.8a (left plot) with its minimum reached at the maximum displayed degree $r^* = 5$. The first three rows of the coefficient matrix $\tilde{A}$ obtained with $r = r^*$ are shown as curves in the bottom plot of fig. 3.7a. These curves correspond to the constant, linear and quadratic components in the polynomial fits $x_j(y)$ of each predictor. The column-wise prediction errors $\chi_j$ are shown in the top plot of fig. 3.7b (solid blue), together with the optimal threshold $\tau_{\mathrm{opt}}$ (horizontal dashed red). The $\rho(\boldsymbol{y})$ error with $r = r^* = 5$ as a function of $\tau$ can be seen in fig. 3.8a (right plot, solid blue), with its minimum indicated by the red square marker. This minimum is close to the minimum of the inaccessible $\rho(\boldsymbol{y}_{\mathrm{t}})$ error (dashed gray). Here, as in the numerical examples of section 3.5, we also observe the synchronous behavior of the feature removal errors on the training and test datasets.

The removed 'improper' predictors are indicated with vertical solid light-gray lines in fig. 3.7b (middle and bottom). We have also extracted a few typical predictor data (colored vertical lines in fig. 3.7a) and displayed them in fig. 3.7b using the same colors and line styles for the corresponding polynomial fits. The retained predictors are in the left plot of fig. 3.7b and the removed predictors are in the right plot. It is easy to recognize the cone-type data, similar to the data in fig. 3.2 (right), in one of the discarded predictors (blue). All predictor data in the discarded gray zone around the dashed blue line in fig. 3.7a have this 'conical' shape. This means that the NIR data in this frequency band is situated on a hyper-cone. Even if this frequency band is not required to build a high-quality predictive IR model, it may become useful for testing other higher-dimensional manifold and nonlinear models.

Finally, in fig. 3.8b we display the standard quality-of-prediction scatter plot which compares the exact values of $\boldsymbol{y}$ against their predictions $\hat{\boldsymbol{y}}$. As can be seen, the removal of the 'improper' predictors significantly improves the quality of prediction. In fact, the prediction error $\rho(\boldsymbol{y}_{\mathrm{t}})$ on the test dataset goes down from 0.28 to 0.05



(a) (b)

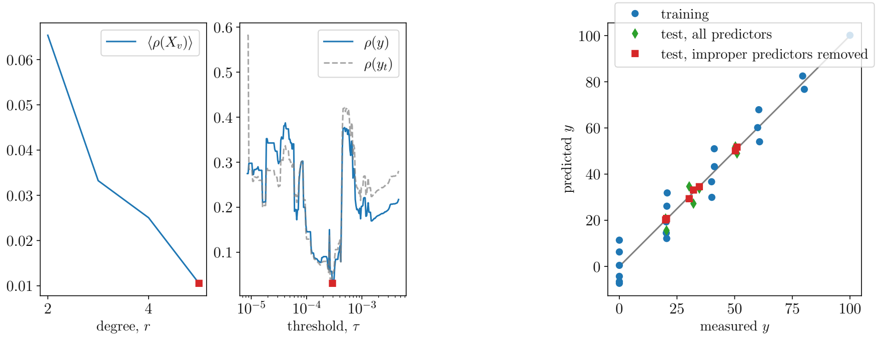
Figure 3.7: (a): column-wise prediction errors $\chi_j$ and the optimal threshold $\tau_{\text{opt}}$ (top); rows of $X$ displayed as curves (middle); first three rows of $\hat{A}$ displayed as curves (bottom). Removed predictors are shown a vertical light-gray lines (middle, bottom). (b): predictor data and the corresponding polynomial fits for the retained (left) and removed (right) predictors.

after the predictor removal procedure.

## 3.7 Summary of results

The hyper-curve PARCUR model introduced in this Chapter provides an alternative, predictor-centered look at the overparameterized multiple linear regression. Within this framework we have been able to focus on the individual roles of predictors and to show that a linear model can be trained and will make perfect predictions even in the presence of predictor variables that violate the linear model assumptions, thus giving an illusion of understanding of the underlying natural phenomena [46]. The column-centered nature of our approach allowed us to come up with a rigorous algorithm for detecting such 'improper' predictors. Moreover, we observe that removing these predictors improves not only the adequacy but also the predictive power of the model.

The polynomial IR version of the PARCUR model has been investigated here in considerable detail. It attempts to build an 'inverse' relation between the dependent variable and each predictor that we believe is much easier to interpret than the weights of the regression vector in the classical MLR model. The polynomial IR model appears to work well with chemometric data, but may also be suitable for other 'smooth' predictors. The main problem with applying the IR model is the difficulty of sorting the dependent variable by magnitude in the presence of noise. While directly using a noisy dependent variable as a parameter in the PARCUR model is not prohibited, it complicates the interpretation of the regression results.



(a)

(b)

Figure 3.8: (a): validation error $\rho(\boldsymbol{y}_{\mathrm{v}})$ as a function of $r$ (left); feature removal error $\rho(\boldsymbol{y})$ at $r = r^*$ as a function of threshold $\tau$ (right). (b): $\hat{\boldsymbol{y}}$ v.s. $\boldsymbol{y}$ for the training dataset (blue circles) and the test dataset predicted at $r = r^*$ with all predictors (green diamonds) and after the removal of 'improper' predictors (red squares).

Certain types of predictor data (microbiome, metabolome, genetic) may exhibit jump-like changes with the dependent variable, especially, in its lower and higher ranges. Successful application of the IR model to these kinds of data will depend on finding a suitable basis for the column function space, e.g., a piecewise-continuous finite-element basis.

Although the PARCUR model allows for a flexible per-column regularization, we have failed to illustrate its potential advantage over the traditional global regularization approach. A separate investigation of this regularization technique is warranted, since it may provide with a more rigorous way to detect and remove 'improper' predictors.

Finally, as we have seen in the application of the polynomial IR model to chemometric data, it can detect the subsets of predictors for which a higher-dimensional model is more appropriate. Nonlinear functional relations between the predictors and the dependent variable produce data that are situated on such higher-dimensional manifolds. A two-parameter hyper-surface model would be a natural extension of the present single-parameter hyper-curve model.

# Chapter 4

# Predicting potato-plant vigor with the DGIR method

[1]

In this chapter we develop a version of the Inverse-Regression (IR) method, called Discontinuous-Galerkin IR (DGIR) method, suitable for predictors that are potentially discontinuous functions of the dependent variable. This chapter also presents the results of the FtV project, i.e., conclusions drawn from the three-year experiment about the predictability of the potato-plant vigor from the seed-tuber properties.

## 4.1 Discontinuous-Galerkin IR model

The explanatory and predictive model of the potato-plant vigor in terms of the seed-tuber properties combines diverse datasets with some of them, like the metabolome data, potentially causing the so-called zero inflation, [35]. This happens when a certain metabolite or a chemical element is detected only in a few samples and is absent in the remaining samples. Other datasets, like FTIR and HSI, where each feature is the absorbance at a given frequency, are generally expected to be continuous not only over the frequency range but also across the samples, see the example of Yarn data at the end of Chapter 3. Our preliminary analysis has shown that, in fact, all types of tuber data may exhibit a discontinuous behavior between different varieties, see Figure 4.1. As one of the goals of the FtV project was to investigate the possibility of a single predictive model that could work with any variety (variety-agnostic model), the regression method should be able to handle such discontinuous data.

[1] This chapter is based on:
E. Atza and N. Budko. Predicting potato plant vigor from the seed tuber properties. *Scientific Reports (Under review)*, 2024. URL `https://doi.org/10.48550/arXiv.2410.19875`

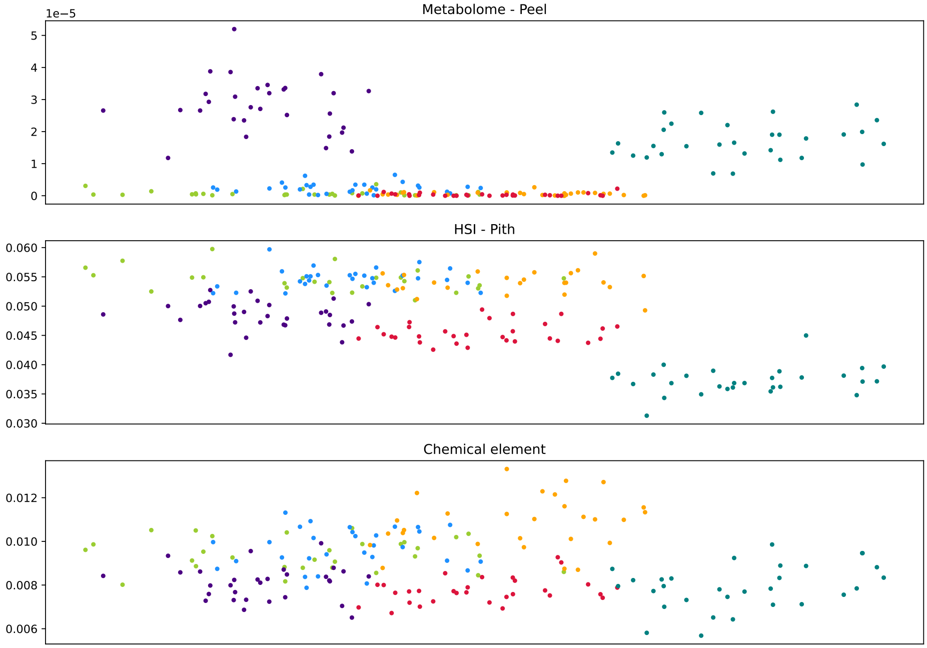


Figure 4.1: Discontinuous behavior of tuber features as functions of plant vigor. The vertical axes correspond to the value of a tuber feature in respective units for metabolome (top), HSI (middle), and XRF (bottom). The horizontal axis corresponds to the plant vigor (magnitude of the canopy in M 2019). The dots are colored according to their variety. The features show variety-specific behaviors and strong discontinuities between some varieties.

Here we use an Inverse Regression (IR) model [4], see Chapter 3, that accommodates for this potential diversity in the predictor data via the freedom to choose a suitable functional basis to represent the predictor variables $x_j$, $j = 1, \dots p$ as functions of the dependent variable $y$. Due to the aforementioned possibility of discontinuities in the predictor data across the samples, we apply the Discontinous-Galerkin Finite-Element Method (DG-FEM) with the first-order (linear) Lagrange basis functions [22] to represent the tuber features as potentially discontinuous functions of plant vigor. Since this is the first time the FEM is applied in the multiple linear regression and possibly also in the general Machine-Learning context, this section elaborates the technical details of the approach.

Let $X \in \mathbb{R}^{n\times p}$ be the predictor data matrix and $\boldsymbol{y} \in \mathbb{R}^n$ – the vector of vigor data sorted by the magnitude of its entries. The idea of the IR method is to represent the predictor data as $x_{i,j} = x_j(y_i)$, i.e., as the (sampled) functions of the dependent variable $y$. These functions are assumed to belong to a finite-dimensional function space defined by the choice of a suitable basis, i.e., a finite-dimensional set of basis functions. With the FTIR data [4] we have applied the monomial basis and a very simple projection procedure leading to the decomposition $X = VA$ of the data matrix in terms of the basis matrix $V$ and the coefficient matrix $A$. A more rigorous and general projection procedure, suitable for a wider class of functions, is the Galerkin projection scheme widely used in the numerical solution of partial differential equations [15].

The derivation of the DG formulation of the IR model starts from the discretization of the $y$-range, i.e., the interval $I = [min(\boldsymbol{y}), max(\boldsymbol{y})]$, into $m$ finite elements $I_k$, $k = 1, \ldots, m$. The number of elements $m$ plays the role of a regularization parameter that will eventually be chosen through a cross validation (CV) procedure, similarly to the maximal polynomial degree in the monomial-basis IR method [4], see Section 3.4, or the maximal dimension of the Krylov subspace in the PLS method [70]. This $y$-range discretization yields $m+1$ delimiters $min(\boldsymbol{y}) = y_0, y_1, \ldots, y_{m-1}, y_m = max(\boldsymbol{y})$. Here, for simplicity, we consider the uniform mesh with the step size $h = y_{k+1} - y_k$.

Next, on each element $I_k = [y_k, y_{k+1}]$, we introduce two local linear Lagrange basis functions $v_{k,1}(y)$ and $v_{k,2}(y)$, defined as:

$$v_{k,1}(y) = \begin{cases} \frac{y_{k+1}-y}{h}, & \text{if } y \in I_k \\ 0, & \text{otherwise} \end{cases}, \qquad v_{k,2}(y) = \begin{cases} \frac{y-y_k}{h}, & \text{if } y \in I_k \\ 0, & \text{otherwise} \end{cases}. \tag{4.1}$$

The main assumption of the IR method is that the predictor data are 'generated' by the functions $x_j(y)$, $j = 1, \ldots, p$; all of which can be represented in terms of the chosen basis as:

$$x_j(y) = \sum_{k=1}^{m} \sum_{q=1}^{2} \alpha_{k,q}^{(j)} v_{k,q}(y), \tag{4.2}$$

and the observed predictor data $x_{i,j}$ are simply the sampled values $x_j(y_i)$, $j = 1, \ldots, p$; $i = 1, \ldots, n$.

The goal of the model training is to determine (learn) the expansion coefficients $\alpha_{k,q}$, $k = 1, \ldots, m$, $q = 1, 2$ in the representation (4.2). For this purpose, we apply the Galerkin scheme, where one chooses the *test* functions to be equal to the basis functions. Multiplying both sides of the equation (4.2) with a test function and integrating over the $y$-range, we arrive at the $2m$ equations for the $2m$ unknown coefficients $\alpha_{k,q}$:

$$\int_{y_0}^{y_m} x_j(y) v_{\hat{k},\hat{q}}(y)\, dy = \sum_{k=1}^{m} \sum_{q=1}^{2} \alpha_{k,q}^{(j)} \int_{y_0}^{y_m} v_{k,q}(y) v_{\hat{k},\hat{q}}(y)\, dy, \qquad \hat{k} = 1, \ldots, m;\ \hat{q} = 1, 2. \tag{4.3}$$

To further specify the system matrix and the right-hand side vector of this linear algebraic system of equations, we first note that, by the nature of the DG finite-element basis, the following partial orthogonality condition holds:

$$\int_{y_0}^{y_m} v_{k,q}(y) v_{\hat{k},\hat{q}}(y)\, dy = \delta_{k,\hat{k}} \int_{y_0}^{y_m} v_{\hat{k},q}(y) v_{\hat{k},\hat{q}}(y)\, dy. \tag{4.4}$$

Therefore, the equations for the coefficients $\alpha_{k,q}$ corresponding to the different elements $I_k$ decouple, and it is possible to find all coefficients by solving $m$ separate $2 \times 2$ linear systems:

$$\int_{y_{\hat{k}-1}}^{y_{\hat{k}}} x_j(y) v_{\hat{k},\hat{q}}(y)\, dy = \sum_{q=1}^{2} \alpha^{(j)}_{\hat{k},q} \int_{y_{\hat{k}-1}}^{y_{\hat{k}}} v_{\hat{k},q}(y) v_{\hat{k},\hat{q}}(y)\, dy, \quad \hat{k} = 1, \dots, m;\ \hat{q} = 1, 2. \tag{4.5}$$

The integrals in (4.5) are further approximated by a numerical quadrature rule [64]. Since the values of $x_j(y)$ are only given at the available data points $y_1, \dots, y_n$, the trapezoidal rule seems to be appropriate:

$$\begin{aligned} \int_{y_{\hat{k}-1}}^{y_{\hat{k}}} v_{\hat{k},q}(y) v_{\hat{k},\hat{q}}(y)\, dy &= \sum_{m=1}^{n_{\hat{k}}} w_{\hat{k},m} v_{\hat{k},q}(y_{\hat{k},m}) v_{\hat{k},\hat{q}}(y_{\hat{k},m}) + \mathcal{O}(\max_m |y_{\hat{k},m+1} - y_{\hat{k},m}|^2), \\ \int_{y_{\hat{k}-1}}^{y_{\hat{k}}} x_j(y) v_{\hat{k},\hat{q}}(y)\, dy &= \sum_{m=1}^{n_{\hat{k}}} w_{\hat{k},m} x_j(y_{\hat{k},m}) v_{\hat{k},\hat{q}}(y_{\hat{k},m}) + \mathcal{O}(\max_m |y_{\hat{k},m+1} - y_{\hat{k},m}|^2), \end{aligned} \tag{4.6}$$

where the data points $y_{\hat{k},m}$, $m = 1, \dots, n_{\hat{k}}$ are situated within the element $I_{\hat{k}}$ bounded by the delimiters $y_{\hat{k}}$ and $y_{\hat{k}+1}$. The weights $w_{\hat{k},m}$ of the trapezoidal quadrature can be explicitly written out as:

$$\begin{aligned} w_{\hat{k},1} &= |y_{\hat{k},1} - y_{\hat{k}}| + \frac{1}{2}|y_{\hat{k},2} - y_{\hat{k},1}|, \\ w_{\hat{k},m} &= \frac{1}{2}|y_{\hat{k},m} - y_{\hat{k},m-1}| + \frac{1}{2}|y_{\hat{k},m+1} - y_{\hat{k},m}|, \quad m = 2, \dots, n_{\hat{k}-1}, \\ w_{\hat{k},n_{\hat{k}}} &= |y_{\hat{k},n_{\hat{k}}} - y_{\hat{k}+1}| + \frac{1}{2}|y_{\hat{k},n_{\hat{k}}} - y_{\hat{k},n_{\hat{k}}-1}|. \end{aligned} \tag{4.7}$$

Therefore we can write a linear approximation of Equation 4.5 for a given interval $I_{\hat{k}}$ with the following notation:

$$V_{\hat{k}} = \begin{bmatrix} v_{\hat{k},1}(y_{\hat{k},1}) & v_{\hat{k},2}(y_{\hat{k},1}) \\ v_{\hat{k},1}(y_{\hat{k},2}) & v_{\hat{k},2}(y_{\hat{k},2}) \\ \vdots & \vdots \\ v_{\hat{k},1}(y_{\hat{k},n_{\hat{k}}}) & v_{\hat{k},2}(y_{\hat{k},n_{\hat{k}}}) \end{bmatrix}, \ W_{\hat{k}} = \operatorname{diag}\left( \begin{bmatrix} w_{\hat{k},1} \\ \vdots \\ w_{\hat{k},n_{\hat{k}}} \end{bmatrix} \right), \ \boldsymbol{x}_j = \begin{bmatrix} x_j(y_{\hat{k},1}) \\ \vdots \\ x_j(y_{\hat{k},n_{\hat{k}}}) \end{bmatrix} \tag{4.8}$$

as

$$V_{\hat{k}}^T W_{\hat{k}} \boldsymbol{x}_j = V_{\hat{k}}^T W_{\hat{k}} V_{\hat{k}} \begin{bmatrix} \alpha_{\hat{k},1}^{(j)} \\ \alpha_{\hat{k},2}^{(j)} \end{bmatrix} = V_{\hat{k}}^T W_{\hat{k}} V_{\hat{k}} \boldsymbol{\alpha}_j \tag{4.9}$$

where the column vector $\boldsymbol{\alpha}_j$ is a subvector of $\boldsymbol{a}_j$, the j-th column of $A$, relative to the $\hat{k}$-th interval.

Since the linear systems concerning each interval are decoupled, the full linear system is:

$$V^T W X = V^T W V A, \quad V^T W \boldsymbol{y} = V^T W V \boldsymbol{a}_0 \tag{4.10}$$

where the block-diagonal matrices and the column-vectors are:

$$V = \begin{bmatrix} V_1 & & \\ & \ddots & \\ & & V_m \end{bmatrix}, \; W = \begin{bmatrix} W_1 & & \\ & \ddots & \\ & & W_m \end{bmatrix}, \; \boldsymbol{a}_j = \begin{bmatrix} \boldsymbol{\alpha}_1^{(j)} \\ \vdots \\ \boldsymbol{\alpha}_m^{(j)} \end{bmatrix}. \tag{4.11}$$

Thus, the coefficient matrix $A$ and the coefficient vector $\boldsymbol{a}_0$ that solve (4.10) can formally be written as:

$$A = (V^T W V)^{-1} V^T W X, \;\; \boldsymbol{a}_0 = (V^T W V)^{-1} V^T W \boldsymbol{y}, \tag{4.12}$$

and are in practice obtained by solving the corresponding linear systems, applying standard parallel-computing techniques for the multiple right-hand sides.

Finally, the prediction $\hat{\boldsymbol{y}}_{\mathrm{t}}$ of the test data-vector $\boldsymbol{y}_{\mathrm{t}}$ has the form

$$\hat{\boldsymbol{y}}_{\mathrm{t}} = \hat{V} \boldsymbol{a}_0, \tag{4.13}$$

where the matrix $\hat{V}$ is obtained by solving the following least-squares problem:

$$\hat{V} = \arg \min_{V \in \mathbb{R}^{s \times 2m}} \|X_{\mathrm{t}} - VA\|_2^2 , \tag{4.14}$$

where $s$ is the number of points in the test set, and $2m$ is the chosen number of basis functions. The matrix $\hat{V}$ is given by:

$$\hat{V} = X_t A^T (AA^T)^{-1}. \tag{4.15}$$

Thus, the prediction $\hat{\boldsymbol{y}}_{\mathrm{t}}$ can be expressed as:

$$\hat{\boldsymbol{y}}_{\mathrm{t}} = X_{\mathrm{t}} A^T (AA^T)^{-1} \boldsymbol{a}_0, \tag{4.16}$$

which again can be obtained by solving the corresponding linear system.

For the purposes of predictor selection, we also compute the prediction $\hat{X}_{\mathrm{t}}$ of the given test data-matrix $X_{\mathrm{t}}$:

$$\hat{X}_{\mathrm{t}} = X_{\mathrm{t}} A^T (AA^T)^{-1} A. \tag{4.17}$$

Computing this prediction involves solving multiple linear systems with the same system matrix and different right-hand sides, which admits efficient parallel implementation.

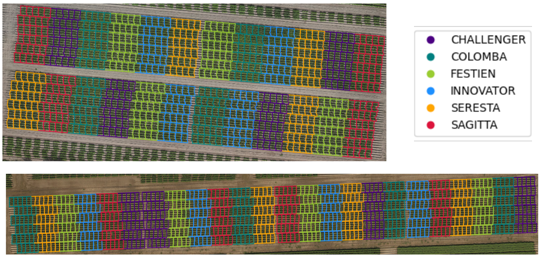


Figure 4.2: Drone images of the V (top) and S (bottom) test fields in 2021 with the overlayed plot boundaries. Colors correspond to the six varieties planted in a randomized block design.

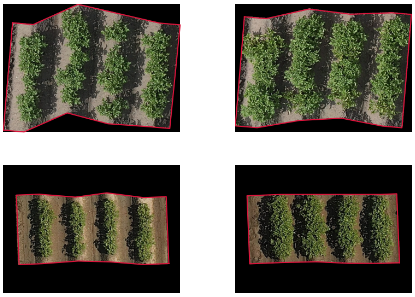


Figure 4.3: Plots of weak (left) and strong (right) seedlots of the Sagitta variety showing consistent performance between the V (top) and the S (bottom) test fields.

## 4.2 Application of the DGIR method

### 4.2.1 Measures of prediction quality

The standard measures of the MLR model quality, such as the $R^2$ coefficient, are not applicable to the severely overparameterized case, where the number of predictors is significantly larger than the number of experiments [37]. To arrive at a similarly interpretable measure, both the measured, $\boldsymbol{y}$, and the predicted, $\hat{\boldsymbol{y}}$, vigor data vectors, each of length $n$, are normalized as:

$$\begin{aligned} \tilde{\boldsymbol{y}} &= \frac{1}{\|\boldsymbol{y} - n^{-1}\mathbf{1}\mathbf{1}^T\boldsymbol{y}\|_2}\left(\boldsymbol{y} - n^{-1}\mathbf{1}\mathbf{1}^T\boldsymbol{y}\right), \\ \tilde{\hat{\boldsymbol{y}}} &= \frac{1}{\|\hat{\boldsymbol{y}} - n^{-1}\mathbf{1}\mathbf{1}^T\hat{\boldsymbol{y}}\|_2}\left(\hat{\boldsymbol{y}} - n^{-1}\mathbf{1}\mathbf{1}^T\hat{\boldsymbol{y}}\right). \end{aligned} \tag{4.18}$$

Here, $\mathbf{1} \in \mathbb{R}^n$ is the vector of all ones. After this transformation, we have $\|\tilde{\boldsymbol{y}}\|_2 = \|\tilde{\hat{\boldsymbol{y}}}\|_2 = 1$, where $\|\cdot\|_2$ denotes the Euclidean norm of the vector, and $n^{-1}\mathbf{1}^T\tilde{\boldsymbol{y}} = n^{-1}\mathbf{1}^T\tilde{\hat{\boldsymbol{y}}} = 0$, i.e., both vectors have the unit norm and the zero mean.

For these normalized vectors we compute the squared norm of the residual vector, also known as Sum of Squared Errors (SSE):

$$r^2 = \|\tilde{\boldsymbol{y}} - \tilde{\hat{\boldsymbol{y}}}\|_2^2, \tag{4.19}$$

which is a measure of the unexplained variance. In this case, the $r^2$ is related to the Pearson correlation coefficient $c(\tilde{\boldsymbol{y}}, \tilde{\hat{\boldsymbol{y}}})$ between the measured and predicted data vectors as follows:

$$r^2 = 2(1 - c). \tag{4.20}$$

Since $-1 \leq c \leq 1$, we have the bounds $0 \leq r^2 \leq 4$. We consider the model to be predictive if for the testing dataset it gives the values $r^2 < 1$ and $c > 0.5$, with the latter parameter also subject to the $p$-value analysis. The upper bound $r^2 = 1$ for the SSE and the corresponding lower bound $c = 0.5$ for the correlation, stem from the fact that for $1 \leq r^2 \leq 4$ the prediction $\tilde{\boldsymbol{y}}$ is usually of very poor quality, making it impossible to correctly classify the vigor of the seedlot as belonging to the low, middle or high quantile.

### 4.2.2 Training, validation and testing

Both the variety-agnostic and the six variety-specific DGIR models contain two hyper-parameters that need to be tuned. The first is the number $m$ of segments, see the Equation (4.2), that divide the range of the vigor parameter. The more segments are used, the better is the fit for each predictor variable. The 10-fold Cross Validation is used to find the optimal number of segments and avoid overfitting. In the DGIR method with linear basis functions, at least two data points should be present within each segment. This puts a practical upper bound on the number of uniform segments, so that one only has to check for a smaller number of segments in search of the optimal parameter $m$. The results presented in Figure 4.6 illustrate the fits with the optimal number of segments determined by the CV procedure.

The second hyper-parameter is the threshold $\tau$ which allows discarding the predictors with the $X$-data residual above $\tau$, see the dashed horizontal lines in Figure 4.5. Tuning of this parameter is performed with the number of segments $m$ fixed at the optimal value.

To tune the abovementioned hyper-parameters of the DGIR model, the dataset is split into the training-validation and testing subsets. When the testing is performed on a different year, then the complete dataset of that year is used as the testing set. This means that the complete dataset of the training year can be used for training and validation (parameter tuning). When, however, the testing is performed on the same year as the training, then 33% of the variety data is reserved for testing and the remaining 67% is used for training and validation.

### 4.2.3 Variety-agnostic modeling

The present experiments were not designed to recover the genotype-by-environment interaction, which was also not known *a priori* for the selected set of varieties. Therefore, the available environmental information (solar radiation, temperature, water etc.) could not be used to its full potential, i.e., in a way that would improve the prediction quality within the specific growth conditions. In these circumstances one can hope to predict the average vigor of a seedlot over several conditions, i.e., over several fields within the same test year. However, this strategy only works if the conditions are not too different between the fields. Since the V field in 2020 and the M field in 2021 did suffer from adverse growth conditions around the emergence times, these fields are considered to be statistical outliers and the corresponding data are treated with extra caution.

To predict the vigor of a plant from the seed tuber properties, a model should be trained on both the tuber and the vigor data and then a prediction should be made using only the tuber data with the measured vigor data used for the testing purposes only. Since the present experiment spanned three years and three fields in each year, there are many possibilities to train and to test the model. For completeness, all possible training-testing possibilities have been implemented and the results are shown as tables in Figures A.1–A.13, where the rows correspond to the training dataset and the columns show the outcome of the testing. Note that the block-diagonal portions of the table correspond to the testing on a randomly selected portion of the dataset that was excluded from the training process. This means that the testing within the same year (diagonal blocks) was performed on a smaller dataset than across the years (off-diagonal blocks) and is therefore less reliable.

The results presented in Figure A.1 correspond to the variety-agnostic model which does not code explicitly for the known variety of the seedlot, and both the SSE residual or the unexplained variance (top table) and the correlation of the predicted and measured vigor (bottom table) are shown. It turns out that the data from 2020 are most suited for the training of the linear model as it results in the best predictions (smallest residuals and highest correlations) for the years 2019 and 2020. Hence, the variety-agnostic model is able to explain around 70% of the variation of the vigor of the plants (within the same year) and to predict from 50% to 70% of the variation in another year. These residuals correspond to 70%-80% correlations between the predicted and measured vitalities, which agrees with the magnitude of the observed correlations (see Figure 2.15, rows marked ’All’) in vigor across the fields in each year.

However, these promising regression results mask the fact that the average vigor (canopy size) of the six considered varieties is markedly different, with the canopy of the ‘weakest’ variety Festien (around $-1$ on the normalized field-centered scale) being almost three times smaller than that of the ‘strongest’ variety Colomba (around 2 on the normalized field-centered scale). Therefore, mathematically speaking, it is easy to achieve a relatively small residual by simply predicting the average vigor of each variety rather than the true vigor of the seedlot.

The fact that the genotype (variety) effect dominates the regression results can be deduced from the scatter-plots of Figure 4.4, where the vigor prediction of each seedlot is colored according to the corresponding genotype. It is obvious that, with some exceptions, the quality of predictions of the seedlot vigor within each individual variety is rather poor.

### 4.2.4 Variety-specific modeling

Since in practice a single variety is usually planted in the field by potato farmers, it is important to know how well one can predict the vigor of a seedlot not across all varieties but within its specific variety. The assumption behind the variety-agnostic model is that the biochemical mechanisms behind a weak or a strong vigor are the same for all potato varieties. However, it is also possible that these mechanisms depend on the variety. Therefore, six separate variety-specific IR models have also been trained and the results are presented in Figures A.2–A.13 next to the corresponding segregated results of the variety-agnostic model.

From Figures A.2–A.13, it is clear that the quality of predictions within the varieties is much lower than across the whole set of varieties, Figure A.1, which confirms the strong genotype effect on the vigor of seedlots. Apart from the statistically less reliable within-the-year results (diagonal blocks in the tables), for the four of the six varieties (Challenger, Innovator, Seresta, Colomba) neither the variety-agnostic nor the variety-specific models are able to explain the intra-varietal variations in the vigor when presented with the tuber data of a different year. Although, the correlations between the predicted and the measured vitalities do fall just a little short of 50% for these varieties, this corresponds to the residuals with the magnitude above one, i.e., such predictions are practically useless.

The notable exception is the Festien variety, which exhibits a fairly predictable behaviour, see Figures A.6, A.7, with the variety-specific model showing somewhat better performance if compared to the variety-agnostic model. The variety-specific Festien model trained on the tuber data of the year 2020 and the average vigor data of the M and S fields in 2020 predicts/explains 30% of the variation of the three-field average vigor in the year 2019 and 27% of the three-field average vigor in 2021, when presented with the corresponding seed tuber data.

In the second place is the Sagitta variety, which shows a similar partially predictable behaviour between the years 2020 and 2021, see Figures A.10–A.11. However, this predictability does not extend to the year 2019 for the Sagitta variety.

### 4.2.5 Predictive power of different tuber data types

The IR model provides a measure of the usefulness of each particular predictor variable [4]. When a new predictor-variable dataset (tuber properties) is considered for prediction purposes, the IR method can perform a 'prediction' of this known dataset by the model created on the training dataset of tuber properties. For example, the dataset of tuber properties in the year 2021 can be predicted from the dataset of tuber properties in the year 2020. The quality of the $X$-data prediction is measured by the normalized residuals for each individual predictor variable, i.e., abundance of a metabolite, spectral amplitude of the FTIR or HSI signature, abundance of an inorganic element. The smallness of such residuals is the necessary condition for the usefulness of the particular predictor variable. One can only achieve a good prediction of the dependent variable (vigor) if these residuals are small.

The predictor-variable residuals, or the $X$-data residuals, can be used in two ways: to rank the predictive power of the different data types, and to perform feature selection.

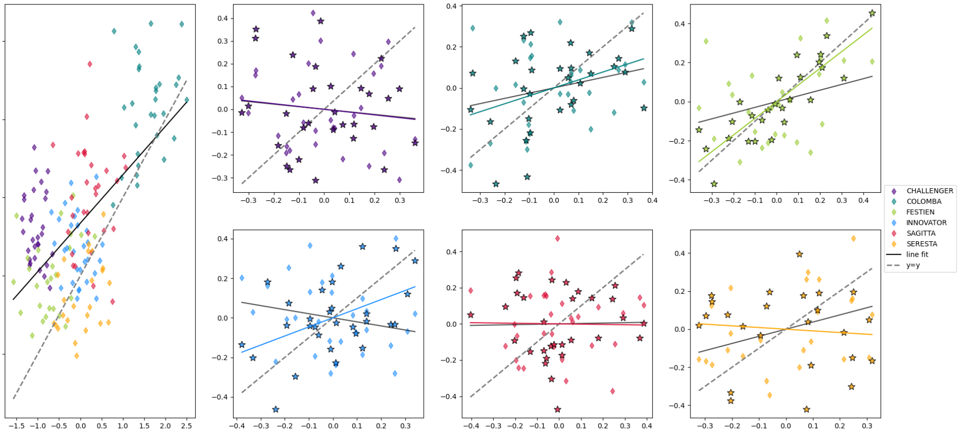


Figure 4.4: Performance of the variety-agnostic and variety-specific models trained on the dataset of the field S in 2021 and tested on the average vigor data of the fields M, S, and V in 2019. Horizontal axis – measured seedlot vigor $y$, vertical axis – predicted seedlot vigor $\hat{y}$. Leftmost vertical panel: variety-agnostic model with the seedlot results colored according to their variety. Dashed gray: the $y = \hat{y}$ line; solid black: the best linear fit between the measured and predicted data. Six panels on the right: prediction per variety with the variety-agnostic model (diamonds) and the variety-specific model (stars) with the variety vigor data (both measured and predicted) shifted and scaled to have zero mean and unit norm. Dashed gray: the $y = \hat{y}$ line; solid black: the best linear fit for the variety agnostic model; solid colored: the best linear fit for the variety-specific model.

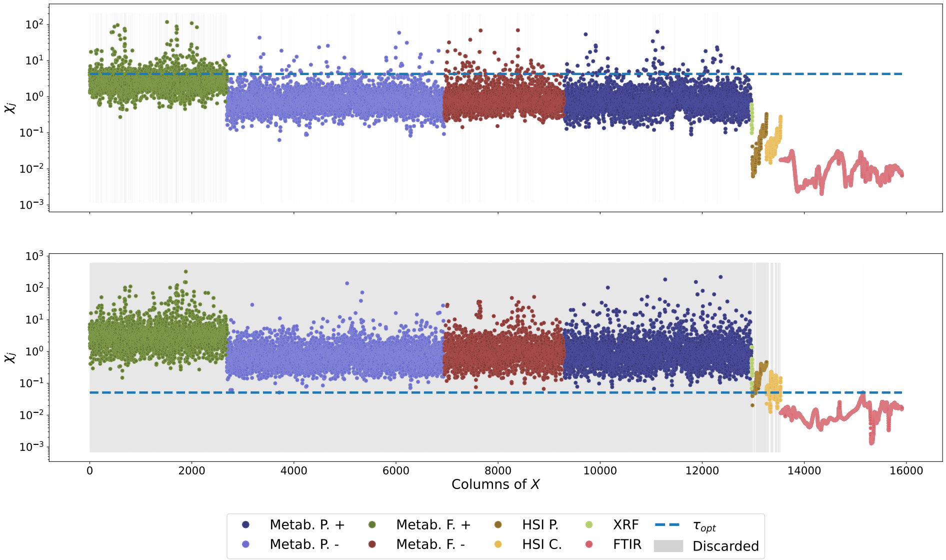


Figure 4.5: The $X$-data residuals for the variety-agnostic (top) and variety-specific (bottom) 2020 models tested on the 2021 Festien tuber data. Colors indicate the data type (see legend). Dashed horizontal lines show the feature selection thresholds, with the rejected predictors marked by vertical gray lines.

Figure 4.5 shows the $X$-data residuals grouped and coloured by the type of the tuber data for the training dataset of the year 2020 and the testing dataset of the year 2021. In the top plot, the residuals of the 2020 variety-agnostic model predicting the 2021 Festien $X$-data are shown, and in the bottom plot, the residuals for the prediction of the same $X$-data by the 2020 variety-specific Festien model are displayed. In this and other training-testing configurations, the FTIR dataset consistently shows the smallest residuals, followed by the two HSI datasets, the XRF and the metabolome datasets.

The metabolome data, on average, demonstrate higher residuals, with some metabolites producing smaller residuals that are still higher than the residuals of the FTIR data. In mathematical terms, this means that the training metabolome dataset is less complete than the training FTIR dataset [4]. There can be several reasons why the IR algorithm indicates such incompleteness. First, the dependence of the majority of metabolites on the vigor parameter is of 'non-functional' type, i.e., does not look like a sampled graph of a function. This is indeed what is observed in the trained variety-agnostic IR model. In Figure 4.6 one can see the fits produced by the IR algorithm for the various predictor variables. The horizontal axis is the vigor parameter, and the vertical axis is the value of the corresponding predictor. The data points are colored in accordance with the seed tuber variety. The metabolome predictors shown

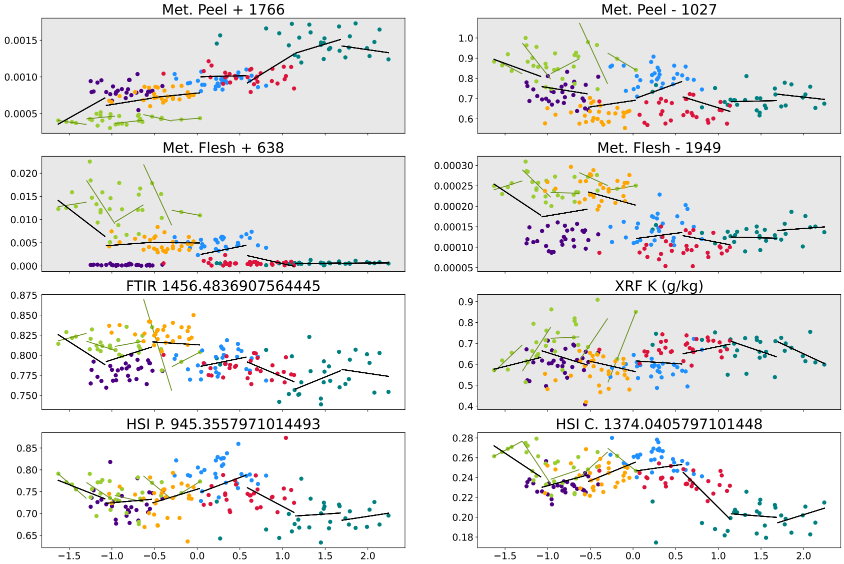


Figure 4.6: Individual predictor data (vertical axes) fitted by piece-wise linear functions of the vigor parameter (horizontal axis). For each data type we show the predictor with the lowest residual in the Festien specific model. The training was performed on the average vigor of the three test fields and the resulting fits by the variety-agnostic model are shown as broken solid black lines. The fits produced by the variety-specific Festien model are shown as broken light-green lines. The data type is indicated in the titles of the plots together with the predictor label. Plots with the light-gray background show predictors that have been rejected by the variety-specific Festien model.

in Figure 4.6 (top four plots) exhibits a typical nonfunctional behavior [4], since the data points from different varieties represent vertically shifted clusters, i.e., it is impossible to draw a graph of a mathematical function of the vigor variable through these clusters.

The second reason for the apparent incompleteness of the metabolome dataset, that can be observed in the variety-specific Festien model, is the relatively high noise in the data, where the data points appear to form a cloud rather than being aligned along a graph of a function, see Figure 4.6 (left column, second plot from the top, green point cloud). The noise in these plots comes from two sources: the vigor data (error in the point location along the horizontal axis) and the metabolome data (error in the point location along the vertical axis). It is not clear which of the two noise sources is dominant, however, it is clear that the noise in the metabolome data is higher than in, e.g., the FTIR or HSI data.

## 4.3 Summary of results

In summary, the variety agnostic model could be a useful tool to identify a range of vitality for a variety but is not very informative when trying to compare different seedlots within a variety. It appears that predicting vigor for seedlots within a variety is not an easily generalizable task, indeed variety specific models perform quite differently. Therefore we cannot achieve the hoped for result of a variety-agnostic model to predict seedlot vitality which generalizes to varieties not included in this project.

The vigor of two varieties, Festien and to a lesser extent Sagitta, appears to be fairly predictable. Although, the model for the Sagitta variety fails to predict the vigor in one of the three test years. Notably, it was also the year which showed relatively small correlations in vigor between the test fields for this variety. Festien and Sagitta are also the better predictable varieties in our microbiome study [58], where it was shown that this level of predictability is enough to distinguish between the three practically meaningful classes for the seedlots of these varieties: below average, average, and above average. Moreover, the feature selection algorithm applied to all biochemical tuber data, indicates that it is sufficient to measure the FTIR spectra of the dry samples of the seed tubers to achieve a residual similar in magnitude to those achieved on the microbiome dataset.

For the other four varieties (Challenger, Colomba, Innovator, and Seresta) the vigor of the seedlots could not be predicted in any reliable way. It is worth noting that the vigor correlations across the test fields were also not as consistent for these varieties, which may be a sign of a purely stochastic nature of vigor variations within these varieties. Nevertheless, in our opinion, the reasons behind this lack of predictability are not entirely clear. On the one hand, it is possible that the biochemical composition of the seed tubers simply does not define the vigor of these varieties to the same extent as with the Festien variety, and there are other processes at play that were not measured in the present experiment. A seed tuber is a slowly changing dynamical system. Therefore, measuring its biochemical constitution at a few points in time may deliver better predictors of the potato plant vigor. This could mean that the tubers of Festien are somehow more 'stationary' than the tubers of other varieties.

# Conclusions

In this thesis a variety of tools has been applied to answer the main research question of the "Flight to Vitality" project: is it possible to predict the vigor of potato plants from biotic and abiotic features of seed tubers?

Having analyzed the data from three years of experiments carried out at different locations in Europe, we have confirmed the hypothesis of practitioners that there exists a significant link between the seed tuber and the plant vigor and have trained Machine-Learning (ML) models that make predictions of the vigor from the tuber properties. The most general variety-agnostic model, trained on the six varieties studied in the FtV project, makes fairly good predictions when tested on the seedlots of these varieties. However, a closer look at the results shows the dominant effect of the genotype on the seedlot vigor and differences in the vigor-enhancing strategies of the varieties. The six trained variety-specific ML models revealed that a more precise prediction of the seedlot vigor is only possible for one of the studied varieties, whereas, for the remaining five varieties the connection between the tuber and the plant vigor is either inconsistent across environments or simply much weaker.

Conducting field experiments is an endeavor subject to uncontrollable but measurable events. In our experiments we registered a strong influence of the environment on the growth of some genotypes and seedlots. Therefore, a closer monitoring of weather and modeling of the genotype-specific interactions with the environment appear to be essential for increasing the accuracy of predictions at the seedlot level. Whether such extended and accurate variety-specific models can be combined into a general model able to predict the vigor of any, even completely new, variety, remains an open question.

These conclusions are heavily dependent on the vigor parameter chosen in this study. It is possible that a model trained to predict a trait other than a "snapshot" of the canopy area, might indeed have a superior performance.

Execution of the FtV project, pre- and post-processing of the data, and the development of predictive ML models required the application of various mathematical methods. Some of these techniques were advanced but reasonably well-known, other were new but theoretically trivial. One mathematical question, however, turned out to be both practically important and not yet completely understood on the theoretical level. We were confronted with this question while developing predictive ML models for the project and pondering about the adequacy and meaning of the standard multiple linear regression (MLR) algorithms, such as the PLS method.

When to switch from a linear to a non-linear model in practical applications where increasing the number of experiments is disproportionally more expensive than increasing the number of predictors? This has lead to the second research question of the thesis and resulted in a fresh perspective on the overparameterized MLR and its reformulation as a parametric hyper-curve (PARCUR) fitting problem. In the inverse-regression (IR) version of PARCUR, we allow the user to adapt the MLR model according to either *a priori* or observed information about the "shape" of predictors as functions of the dependent variable. The IR method was developed for continuous functions and extended to discontinuous functions with the help of the Discontinuous-Galerkin projection procedure when applying the model to the FtV data.

On the theoretical level we prove that an overparameterized MLR model will be able to make exact predictions even in the presence of a nonlinear relation between the predictors and the dependent variable. Hence, from the practical point of view, there may be no need for a nonlinear model even in the case of a (partially) nonlinear relation. Since this fact may lead to an illusion of understanding, it is important to identify and remove such nonlinear relations from the MLR model. We show that the previously unutilized step of projecting the new but known predictor data on the learned representation of predictors provides an efficient and objective way of detecting such nonlinear relations. Moreover, this procedure also helps to identify and exclude the noisy predictor data, thereby improving not only the adequacy of the model but also the accuracy of its predictions.

Returning to the FtV project and its experiments, we regret not having been able to use the climate-controlled room data and the multi-spectral images in the fields to their full potential due to the problems mentioned in Chapter 1. We believe that the rapid progress in high-throughput phenotyping and the improving quality of the data-acquisition and pre-processing techniques will make these types of data more accessible in the future projects.

As far the development of new mathematical methods for agricultural applications is concerned, we would like to point out the open nature of the field-effect removal problem. While the state-of-the-art software employed in this project allows removing a smooth spatial trend, it does not elucidate the underlying physical and biological causes of the test-field heterogeneity. A better model should connect the genotype-specific physiological model of plant reaction to various environmental factors with the physical model and local measurements of the corresponding environmental variables.

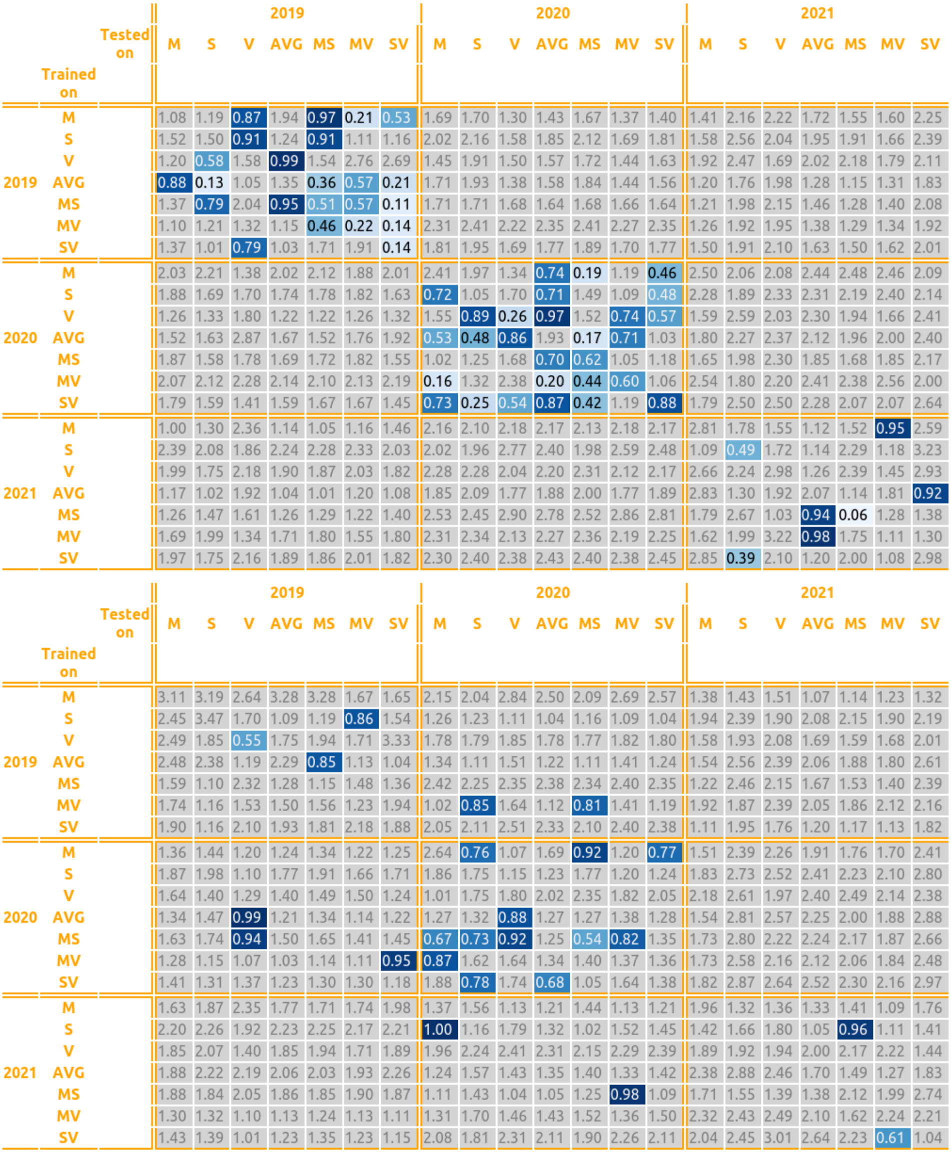
# Appendix A

# Figures

| | Tested on | 2019 | | | | | | | 2020 | | | | | | | 2021 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Trained on | | M | S | V | AVG | MS | MV | SV | M | S | V | AVG | MS | MV | SV | M | S | V | AVG | MS | MV | SV |
| 2019 | M | 0.56 | 0.88 | 0.73 | 0.62 | 0.61 | 0.47 | 0.38 | 1.04 | 0.83 | 0.98 | 0.88 | 0.89 | 0.95 | 0.86 | 1.77 | 1.22 | 1.12 | 1.09 | 1.25 | 1.17 | 1.14 |
| | S | 0.61 | 0.42 | 0.46 | 0.42 | 0.34 | 0.64 | 0.35 | 0.63 | 0.54 | 0.65 | 0.51 | 0.52 | 0.55 | 0.54 | 2.27 | 0.57 | 0.44 | 0.69 | 1.14 | 1.03 | 0.45 |
| | V | 0.51 | 0.98 | 0.88 | 0.81 | 0.83 | 0.84 | 0.64 | 0.92 | 0.79 | 0.99 | 0.83 | 0.80 | 0.89 | 0.85 | 1.83 | 1.20 | 1.08 | 1.09 | 1.28 | 1.19 | 1.11 |
| | AVG | 0.50 | 0.71 | 0.55 | 0.48 | 0.43 | 0.47 | 0.53 | 0.85 | 0.70 | 0.87 | 0.73 | 0.72 | 0.79 | 0.74 | 1.70 | 0.75 | 0.97 | 0.76 | 0.85 | 1.01 | 0.82 |
| | MS | 0.66 | 0.62 | 0.83 | 0.52 | 0.42 | 0.77 | 0.46 | 1.26 | 0.83 | 1.01 | 0.97 | 1.00 | 1.09 | 0.88 | 1.85 | 1.56 | 1.65 | 1.54 | 1.56 | 1.62 | 1.59 |
| | MV | 0.56 | 0.98 | 0.49 | 0.58 | 0.68 | 0.20 | 0.44 | 0.58 | 0.56 | 0.58 | 0.48 | 0.50 | 0.50 | 0.52 | 1.75 | 0.58 | 0.54 | 0.50 | 0.76 | 0.72 | 0.51 |
| | SV | 0.75 | 0.42 | 0.73 | 0.32 | 0.28 | 0.44 | 0.32 | 1.06 | 0.87 | 0.84 | 0.86 | 0.92 | 0.89 | 0.81 | 1.57 | 1.25 | 1.17 | 1.03 | 1.12 | 1.06 | 1.18 |
| 2020 | M | 0.64 | 0.69 | 0.64 | 0.55 | 0.56 | 0.60 | 0.56 | 0.24 | 0.35 | 0.45 | 0.31 | 0.27 | 0.31 | 0.35 | 1.53 | 0.50 | 0.47 | 0.31 | 0.53 | 0.50 | 0.43 |
| | S | 0.57 | 0.63 | 0.60 | 0.49 | 0.49 | 0.55 | 0.51 | 0.30 | 0.15 | 0.35 | 0.39 | 0.14 | 0.53 | 0.36 | 1.94 | 0.41 | 0.28 | 0.38 | 0.77 | 0.67 | 0.28 |
| | V | 1.03 | 0.81 | 1.03 | 0.87 | 0.84 | 1.01 | 0.84 | 0.38 | 0.44 | 0.61 | 0.51 | 0.52 | 0.56 | 0.51 | 1.80 | 0.51 | 0.51 | 0.47 | 0.74 | 0.74 | 0.45 |
| | AVG | 0.69 | 0.57 | 0.76 | 0.57 | 0.52 | 0.69 | 0.56 | 0.34 | 0.39 | 0.50 | 0.23 | 0.30 | 0.24 | 0.32 | 1.67 | 0.44 | 0.33 | 0.29 | 0.59 | 0.50 | 0.32 |
| | MS | 0.79 | 0.75 | 0.83 | 0.70 | 0.68 | 0.78 | 0.70 | 0.41 | 0.27 | 0.49 | 0.27 | 0.26 | 0.36 | 0.34 | 1.86 | 0.46 | 0.38 | 0.41 | 0.75 | 0.68 | 0.36 |
| | MV | 0.74 | 0.67 | 0.81 | 0.64 | 0.61 | 0.74 | 0.65 | 0.44 | 0.35 | 0.46 | 0.31 | 0.35 | 0.32 | 0.37 | 1.82 | 0.38 | 0.34 | 0.34 | 0.66 | 0.63 | 0.30 |
| | SV | 0.62 | 0.40 | 0.61 | 0.42 | 0.39 | 0.58 | 0.39 | 0.54 | 0.26 | 0.69 | 0.36 | 0.32 | 0.49 | 0.37 | 1.73 | 0.49 | 0.45 | 0.40 | 0.67 | 0.64 | 0.41 |
| 2021 | M | 1.65 | 1.90 | 1.78 | 1.76 | 1.76 | 1.71 | 1.83 | 1.46 | 1.99 | 1.71 | 1.70 | 1.71 | 1.56 | 1.85 | 0.39 | 2.48 | 2.20 | 1.60 | 1.21 | 1.06 | 2.28 |
| | S | 0.89 | 0.89 | 0.82 | 0.78 | 0.81 | 0.83 | 0.77 | 0.92 | 0.71 | 1.21 | 0.88 | 0.77 | 1.00 | 0.92 | 2.51 | 0.34 | 0.45 | 0.73 | 1.17 | 0.66 | 0.28 |
| | V | 0.63 | 0.66 | 0.67 | 0.54 | 0.54 | 0.61 | 0.56 | 1.11 | 0.77 | 0.88 | 0.85 | 0.89 | 0.94 | 0.78 | 2.01 | 0.40 | 0.27 | 0.37 | 0.81 | 0.70 | 0.24 |
| | AVG | 1.16 | 1.35 | 1.07 | 1.13 | 1.20 | 1.09 | 1.15 | 0.98 | 1.51 | 1.38 | 1.24 | 1.21 | 1.12 | 1.42 | 1.25 | 0.53 | 0.65 | 0.38 | 0.45 | 0.58 | 0.53 |
| | MS | 1.93 | 2.05 | 1.91 | 1.96 | 1.99 | 1.92 | 1.98 | 0.86 | 1.54 | 1.30 | 1.18 | 1.16 | 1.02 | 1.40 | 1.23 | 0.81 | 1.08 | 0.61 | 0.67 | 0.71 | 1.22 |
| | MV | 1.13 | 1.61 | 1.25 | 1.28 | 1.33 | 1.17 | 1.39 | 0.85 | 1.41 | 1.33 | 1.14 | 1.08 | 1.03 | 1.35 | 1.28 | 1.09 | 0.87 | 0.57 | 0.60 | 0.41 | 0.96 |
| | SV | 1.24 | 0.92 | 1.34 | 1.10 | 1.01 | 1.27 | 1.06 | 0.51 | 0.95 | 0.85 | 0.69 | 0.67 | 0.59 | 0.86 | 2.28 | 0.41 | 0.32 | 0.57 | 1.02 | 0.97 | 0.30 |

| | Tested on | 2019 | | | | | | | 2020 | | | | | | | 2021 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Trained on | | M | S | V | AVG | MS | MV | SV | M | S | V | AVG | MS | MV | SV | M | S | V | AVG | MS | MV | SV |
| 2019 | M | 0.72 | 0.56 | 0.63 | 0.69 | 0.70 | 0.77 | 0.81 | 0.48 | 0.59 | 0.51 | 0.56 | 0.56 | 0.52 | 0.57 | | 0.39 | 0.44 | 0.45 | 0.37 | 0.42 | 0.43 |
| | S | 0.70 | 0.79 | 0.77 | 0.79 | 0.83 | 0.68 | 0.82 | 0.69 | 0.73 | 0.68 | 0.74 | 0.74 | 0.72 | 0.73 | | 0.71 | 0.78 | 0.66 | 0.43 | 0.48 | 0.77 |
| | V | 0.75 | 0.51 | 0.56 | 0.60 | 0.59 | 0.58 | 0.68 | 0.54 | 0.61 | 0.50 | 0.59 | 0.60 | 0.55 | 0.58 | | 0.40 | 0.46 | 0.45 | 0.36 | 0.40 | 0.44 |
| | AVG | 0.75 | 0.64 | 0.73 | 0.76 | 0.79 | 0.76 | 0.74 | 0.57 | 0.65 | 0.57 | 0.64 | 0.64 | 0.60 | 0.63 | 0.15 | 0.62 | 0.51 | 0.62 | 0.58 | 0.50 | 0.59 |
| | MS | 0.67 | 0.69 | 0.58 | 0.74 | 0.79 | 0.61 | 0.77 | 0.37 | 0.58 | 0.49 | 0.51 | 0.50 | 0.46 | 0.56 | | 0.22 | 0.18 | 0.23 | 0.22 | 0.19 | 0.21 |
| | MV | 0.72 | 0.51 | 0.75 | 0.71 | 0.66 | 0.90 | 0.78 | 0.71 | 0.72 | 0.71 | 0.76 | 0.75 | 0.75 | 0.74 | | 0.71 | 0.73 | 0.75 | 0.62 | 0.64 | 0.75 |
| | SV | 0.63 | 0.79 | 0.64 | 0.84 | 0.86 | 0.78 | 0.84 | 0.47 | 0.56 | 0.58 | 0.57 | 0.54 | 0.55 | 0.59 | 0.22 | 0.37 | 0.41 | 0.48 | 0.44 | 0.47 | 0.41 |
| 2020 | M | 0.68 | 0.66 | 0.68 | 0.73 | 0.72 | 0.70 | 0.72 | 0.88 | 0.83 | 0.78 | 0.85 | 0.87 | 0.85 | 0.82 | 0.24 | 0.75 | 0.76 | 0.84 | 0.73 | 0.75 | 0.79 |
| | S | 0.72 | 0.68 | 0.70 | 0.76 | 0.75 | 0.72 | 0.74 | 0.85 | 0.93 | 0.82 | 0.80 | 0.93 | 0.74 | 0.82 | | 0.79 | 0.86 | 0.81 | 0.61 | 0.67 | 0.86 |
| | V | 0.48 | 0.59 | 0.48 | 0.56 | 0.58 | 0.50 | 0.58 | 0.81 | 0.78 | 0.70 | 0.75 | 0.74 | 0.72 | 0.75 | | 0.75 | 0.75 | 0.77 | 0.63 | 0.63 | 0.77 |
| | AVG | 0.65 | 0.72 | 0.62 | 0.72 | 0.74 | 0.65 | 0.72 | 0.83 | 0.81 | 0.75 | 0.89 | 0.85 | 0.88 | 0.84 | 0.16 | 0.78 | 0.84 | 0.86 | 0.70 | 0.75 | 0.84 |
| | MS | 0.60 | 0.62 | 0.58 | 0.65 | 0.66 | 0.61 | 0.65 | 0.79 | 0.87 | 0.76 | 0.87 | 0.87 | 0.82 | 0.83 | | 0.77 | 0.81 | 0.79 | 0.62 | 0.66 | 0.82 |
| | MV | 0.63 | 0.66 | 0.60 | 0.68 | 0.70 | 0.63 | 0.68 | 0.78 | 0.82 | 0.77 | 0.85 | 0.83 | 0.84 | 0.81 | | 0.81 | 0.83 | 0.83 | 0.67 | 0.69 | 0.85 |
| | SV | 0.69 | 0.80 | 0.69 | 0.79 | 0.80 | 0.71 | 0.80 | 0.73 | 0.87 | 0.65 | 0.82 | 0.84 | 0.75 | 0.81 | | 0.76 | 0.77 | 0.80 | 0.66 | 0.68 | 0.79 |
| 2021 | M | 0.17 | | | | | | | 0.27 | | | 0.15 | | 0.22 | | 0.80 | | | | 0.40 | 0.47 | |
| | S | 0.56 | 0.55 | 0.59 | 0.61 | 0.60 | 0.59 | 0.61 | 0.54 | 0.64 | 0.40 | 0.56 | 0.62 | 0.50 | 0.54 | | 0.83 | 0.77 | 0.64 | 0.42 | 0.67 | 0.86 |
| | V | 0.69 | 0.67 | 0.67 | 0.73 | 0.73 | 0.69 | 0.72 | 0.44 | 0.62 | 0.56 | 0.57 | 0.55 | 0.53 | 0.61 | | 0.80 | 0.87 | 0.81 | 0.60 | 0.65 | 0.88 |
| | AVG | 0.42 | 0.32 | 0.46 | 0.43 | 0.40 | 0.45 | 0.42 | 0.51 | 0.24 | 0.31 | 0.38 | 0.40 | 0.44 | 0.29 | 0.37 | 0.73 | 0.67 | 0.81 | 0.77 | 0.71 | 0.74 |
| | MS | | | | | | | | 0.57 | 0.23 | 0.35 | 0.41 | 0.42 | 0.49 | 0.30 | 0.39 | 0.60 | 0.46 | 0.70 | 0.67 | 0.65 | 0.39 |
| | MV | 0.43 | 0.19 | 0.38 | 0.36 | 0.34 | 0.41 | 0.31 | 0.58 | 0.30 | 0.33 | 0.43 | 0.46 | 0.49 | 0.33 | 0.36 | 0.45 | 0.56 | 0.72 | 0.70 | 0.80 | 0.52 |
| | SV | 0.38 | 0.54 | 0.33 | 0.45 | 0.50 | 0.36 | 0.47 | 0.75 | 0.52 | 0.58 | 0.66 | 0.67 | 0.70 | 0.57 | | 0.80 | 0.84 | 0.72 | 0.49 | 0.51 | 0.85 |

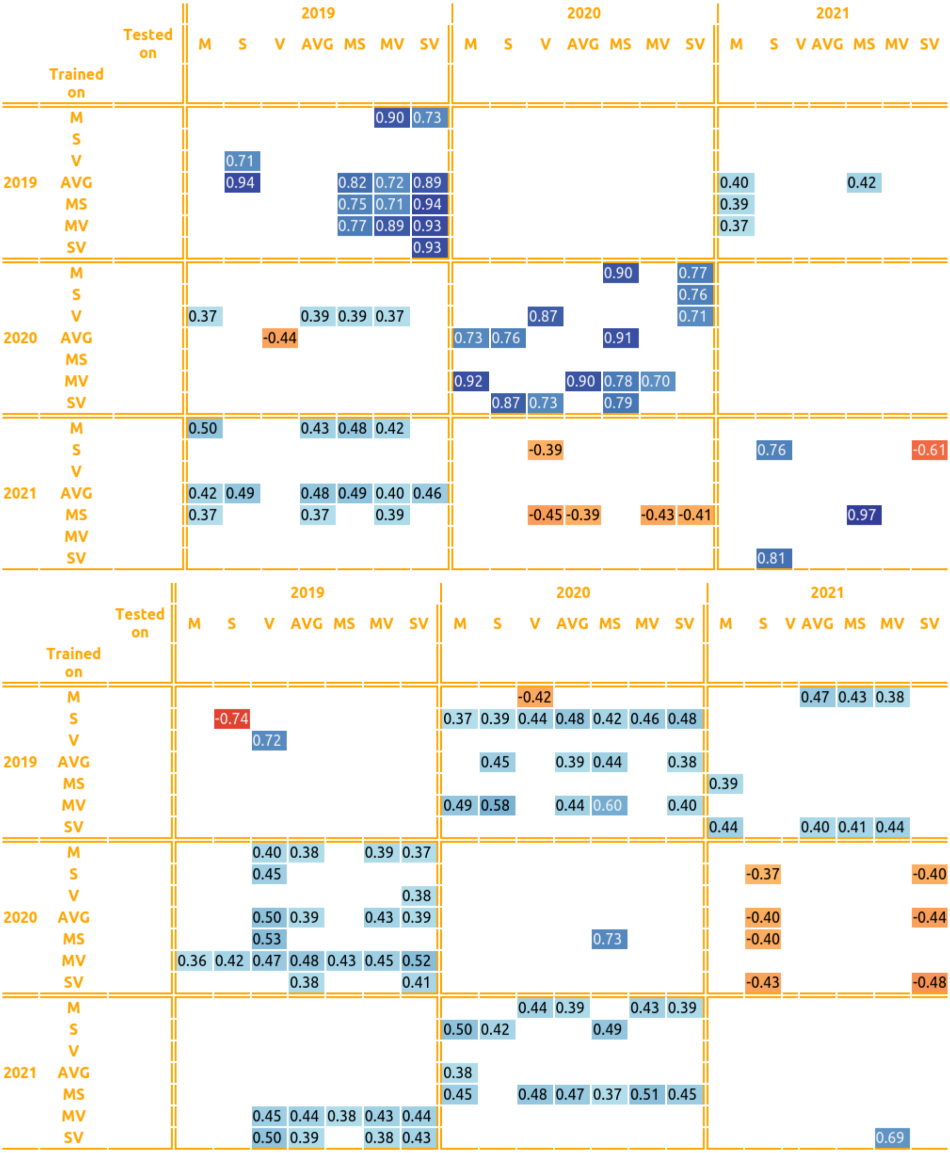
Figure A.1: Performance of the variety-agnostic model tested on all six varieties simultaneously: the SSE residual $r^2$ (top), the Pearson correlation coefficient $c$ (bottom). Rows correspond to the training dataset, columns show the testing configuration. SSE residuals $r^2 \geq 1$ are grayed out, correlations with $p > 0.05$ are omitted.



| Trained on | Tested on | 2019 | | | | | | | 2020 | | | | | | | 2021 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | M | S | V | AVG | MS | MV | SV | M | S | V | AVG | MS | MV | SV | M | S | V | AVG | MS | MV | SV |
| | M | 1.08 | 1.19 | 0.87 | 1.94 | 0.97 | 0.21 | 0.53 | 1.69 | 1.70 | 1.30 | 1.43 | 1.67 | 1.37 | 1.40 | 1.41 | 2.16 | 2.22 | 1.72 | 1.55 | 1.60 | 2.25 |
| | S | 1.52 | 1.50 | 0.91 | 1.24 | 0.91 | 1.11 | 1.16 | 2.02 | 2.16 | 1.58 | 1.85 | 2.12 | 1.69 | 1.81 | 1.58 | 2.56 | 2.04 | 1.95 | 1.91 | 1.66 | 2.39 |
| | V | 1.20 | 0.58 | 1.58 | 0.99 | 1.54 | 2.76 | 2.69 | 1.45 | 1.91 | 1.50 | 1.57 | 1.72 | 1.44 | 1.63 | 1.92 | 2.47 | 1.69 | 2.02 | 2.18 | 1.79 | 2.11 |
| 2019 | AVG | 0.88 | 0.13 | 1.05 | 1.35 | 0.36 | 0.57 | 0.21 | 1.71 | 1.93 | 1.38 | 1.58 | 1.84 | 1.44 | 1.56 | 1.20 | 1.76 | 1.98 | 1.28 | 1.15 | 1.31 | 1.83 |
| | MS | 1.37 | 0.79 | 2.04 | 0.95 | 0.51 | 0.57 | 0.11 | 1.71 | 1.71 | 1.68 | 1.64 | 1.68 | 1.66 | 1.64 | 1.21 | 1.98 | 2.15 | 1.46 | 1.28 | 1.40 | 2.08 |
| | MV | 1.10 | 1.21 | 1.32 | 1.15 | 0.46 | 0.22 | 0.14 | 2.31 | 2.41 | 2.22 | 2.35 | 2.41 | 2.27 | 2.35 | 1.26 | 1.92 | 1.95 | 1.38 | 1.29 | 1.34 | 1.92 |
| | SV | 1.37 | 1.01 | 0.79 | 1.03 | 1.71 | 1.91 | 0.14 | 1.81 | 1.95 | 1.69 | 1.77 | 1.89 | 1.70 | 1.77 | 1.50 | 1.91 | 2.10 | 1.63 | 1.50 | 1.62 | 2.01 |
| | M | 2.03 | 2.21 | 1.38 | 2.02 | 2.12 | 1.88 | 2.01 | 2.41 | 1.97 | 1.34 | 0.74 | 0.19 | 1.19 | 0.46 | 2.50 | 2.06 | 2.08 | 2.44 | 2.48 | 2.46 | 2.09 |
| | S | 1.88 | 1.69 | 1.70 | 1.74 | 1.78 | 1.82 | 1.63 | 0.72 | 1.05 | 1.70 | 0.71 | 1.49 | 1.09 | 0.48 | 2.28 | 1.89 | 2.33 | 2.31 | 2.19 | 2.40 | 2.14 |
| | V | 1.26 | 1.33 | 1.80 | 1.22 | 1.22 | 1.26 | 1.32 | 1.55 | 0.89 | 0.26 | 0.97 | 1.52 | 0.74 | 0.57 | 1.59 | 2.59 | 2.03 | 2.30 | 1.94 | 1.66 | 2.41 |
| 2020 | AVG | 1.52 | 1.63 | 2.87 | 1.67 | 1.52 | 1.76 | 1.92 | 0.53 | 0.48 | 0.86 | 1.93 | 0.17 | 0.71 | 1.03 | 1.80 | 2.27 | 2.37 | 2.12 | 1.96 | 2.00 | 2.40 |
| | MS | 1.87 | 1.58 | 1.78 | 1.69 | 1.72 | 1.82 | 1.55 | 1.02 | 1.25 | 1.68 | 0.70 | 0.62 | 1.05 | 1.18 | 1.65 | 1.98 | 2.30 | 1.85 | 1.68 | 1.85 | 2.17 |
| | MV | 2.07 | 2.12 | 2.28 | 2.14 | 2.10 | 2.13 | 2.19 | 0.16 | 1.32 | 2.38 | 0.20 | 0.44 | 0.60 | 1.06 | 2.54 | 1.80 | 2.20 | 2.41 | 2.38 | 2.56 | 2.00 |
| | SV | 1.79 | 1.59 | 1.41 | 1.59 | 1.67 | 1.67 | 1.45 | 0.73 | 0.25 | 0.54 | 0.87 | 0.42 | 1.19 | 0.88 | 1.79 | 2.50 | 2.50 | 2.28 | 2.07 | 2.07 | 2.64 |
| | M | 1.00 | 1.30 | 2.36 | 1.14 | 1.05 | 1.16 | 1.46 | 2.16 | 2.10 | 2.18 | 2.17 | 2.13 | 2.18 | 2.17 | 2.81 | 1.78 | 1.55 | 1.12 | 1.52 | 0.95 | 2.59 |
| | S | 2.39 | 2.08 | 1.86 | 2.24 | 2.28 | 2.33 | 2.03 | 2.02 | 1.96 | 2.77 | 2.40 | 1.98 | 2.59 | 2.48 | 1.09 | 0.49 | 1.72 | 1.14 | 2.29 | 1.18 | 3.23 |
| | V | 1.99 | 1.75 | 2.18 | 1.90 | 1.87 | 2.03 | 1.82 | 2.28 | 2.28 | 2.04 | 2.20 | 2.31 | 2.12 | 2.17 | 2.66 | 2.24 | 2.98 | 1.26 | 2.39 | 1.45 | 2.93 |
| 2021 | AVG | 1.17 | 1.02 | 1.92 | 1.04 | 1.01 | 1.20 | 1.08 | 1.85 | 2.09 | 1.77 | 1.88 | 2.00 | 1.77 | 1.89 | 2.83 | 1.30 | 1.92 | 2.07 | 1.14 | 1.81 | 0.92 |
| | MS | 1.26 | 1.47 | 1.61 | 1.26 | 1.29 | 1.22 | 1.40 | 2.53 | 2.45 | 2.90 | 2.78 | 2.52 | 2.86 | 2.81 | 1.79 | 2.67 | 1.03 | 0.94 | 0.06 | 1.28 | 1.38 |
| | MV | 1.69 | 1.99 | 1.34 | 1.71 | 1.80 | 1.55 | 1.80 | 2.31 | 2.34 | 2.13 | 2.27 | 2.36 | 2.19 | 2.25 | 1.62 | 1.99 | 3.22 | 0.98 | 1.75 | 1.11 | 1.30 |
| | SV | 1.97 | 1.75 | 2.16 | 1.89 | 1.86 | 2.01 | 1.82 | 2.30 | 2.40 | 2.38 | 2.43 | 2.40 | 2.38 | 2.45 | 2.85 | 0.39 | 2.10 | 1.20 | 2.00 | 1.08 | 2.98 |

| Trained on | Tested on | 2019 | | | | | | | 2020 | | | | | | | 2021 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | M | S | V | AVG | MS | MV | SV | M | S | V | AVG | MS | MV | SV | M | S | V | AVG | MS | MV | SV |
| | M | 3.11 | 3.19 | 2.64 | 3.28 | 3.28 | 1.67 | 1.65 | 2.15 | 2.04 | 2.84 | 2.50 | 2.09 | 2.69 | 2.57 | 1.38 | 1.43 | 1.51 | 1.07 | 1.14 | 1.23 | 1.32 |
| | S | 2.45 | 3.47 | 1.70 | 1.09 | 1.19 | 0.86 | 1.54 | 1.26 | 1.23 | 1.11 | 1.04 | 1.16 | 1.09 | 1.04 | 1.94 | 2.39 | 1.90 | 2.08 | 2.15 | 1.90 | 2.19 |
| | V | 2.49 | 1.85 | 0.55 | 1.75 | 1.94 | 1.71 | 3.33 | 1.78 | 1.79 | 1.85 | 1.78 | 1.77 | 1.82 | 1.80 | 1.58 | 1.93 | 2.08 | 1.69 | 1.59 | 1.68 | 2.01 |
| 2019 | AVG | 2.48 | 2.38 | 1.19 | 2.29 | 0.85 | 1.13 | 1.04 | 1.34 | 1.11 | 1.51 | 1.22 | 1.11 | 1.41 | 1.24 | 1.54 | 2.56 | 2.39 | 2.06 | 1.88 | 1.80 | 2.61 |
| | MS | 1.59 | 1.10 | 2.32 | 1.28 | 1.15 | 1.48 | 1.36 | 2.42 | 2.25 | 2.35 | 2.38 | 2.34 | 2.40 | 2.35 | 1.22 | 2.46 | 2.15 | 1.67 | 1.53 | 1.40 | 2.39 |
| | MV | 1.74 | 1.16 | 1.53 | 1.50 | 1.56 | 1.23 | 1.94 | 1.02 | 0.85 | 1.64 | 1.12 | 0.81 | 1.41 | 1.19 | 1.92 | 1.87 | 2.39 | 2.05 | 1.86 | 2.12 | 2.16 |
| | SV | 1.90 | 1.16 | 2.10 | 1.93 | 1.81 | 2.18 | 1.88 | 2.05 | 2.11 | 2.51 | 2.33 | 2.10 | 2.40 | 2.38 | 1.11 | 1.95 | 1.76 | 1.20 | 1.17 | 1.13 | 1.82 |
| | M | 1.36 | 1.44 | 1.20 | 1.24 | 1.34 | 1.22 | 1.25 | 2.64 | 0.76 | 1.07 | 1.69 | 0.92 | 1.20 | 0.77 | 1.51 | 2.39 | 2.26 | 1.91 | 1.76 | 1.70 | 2.41 |
| | S | 1.87 | 1.98 | 1.10 | 1.77 | 1.91 | 1.66 | 1.71 | 1.86 | 1.75 | 1.15 | 1.23 | 1.77 | 1.20 | 1.24 | 1.83 | 2.73 | 2.52 | 2.41 | 2.23 | 2.10 | 2.80 |
| | V | 1.64 | 1.40 | 1.29 | 1.40 | 1.49 | 1.50 | 1.24 | 1.01 | 1.75 | 1.80 | 2.02 | 2.35 | 1.82 | 2.05 | 2.18 | 2.61 | 1.97 | 2.40 | 2.49 | 2.14 | 2.38 |
| 2020 | AVG | 1.34 | 1.47 | 0.99 | 1.21 | 1.34 | 1.14 | 1.22 | 1.27 | 1.32 | 0.88 | 1.27 | 1.27 | 1.38 | 1.28 | 1.54 | 2.81 | 2.57 | 2.25 | 2.00 | 1.88 | 2.88 |
| | MS | 1.63 | 1.74 | 0.94 | 1.50 | 1.65 | 1.41 | 1.45 | 0.67 | 0.73 | 0.92 | 1.25 | 0.54 | 0.82 | 1.35 | 1.73 | 2.80 | 2.22 | 2.24 | 2.17 | 1.87 | 2.66 |
| | MV | 1.28 | 1.15 | 1.07 | 1.03 | 1.14 | 1.11 | 0.95 | 0.87 | 1.62 | 1.64 | 1.34 | 1.40 | 1.37 | 1.36 | 1.73 | 2.58 | 2.16 | 2.12 | 2.06 | 1.84 | 2.48 |
| | SV | 1.41 | 1.31 | 1.37 | 1.23 | 1.30 | 1.30 | 1.18 | 1.88 | 0.78 | 1.74 | 0.68 | 1.05 | 1.64 | 1.38 | 1.82 | 2.87 | 2.64 | 2.52 | 2.30 | 2.16 | 2.97 |
| | M | 1.63 | 1.87 | 2.35 | 1.77 | 1.71 | 1.74 | 1.98 | 1.37 | 1.56 | 1.13 | 1.21 | 1.44 | 1.13 | 1.21 | 1.96 | 1.32 | 1.36 | 1.33 | 1.41 | 1.09 | 1.76 |
| | S | 2.20 | 2.26 | 1.92 | 2.23 | 2.25 | 2.17 | 2.21 | 1.00 | 1.16 | 1.79 | 1.32 | 1.02 | 1.52 | 1.45 | 1.42 | 1.66 | 1.80 | 1.05 | 0.96 | 1.11 | 1.41 |
| | V | 1.85 | 2.07 | 1.40 | 1.85 | 1.94 | 1.71 | 1.89 | 1.96 | 2.24 | 2.41 | 2.31 | 2.15 | 2.29 | 2.39 | 1.89 | 1.92 | 1.94 | 2.00 | 2.17 | 2.22 | 1.44 |
| 2021 | AVG | 1.88 | 2.22 | 2.19 | 2.06 | 2.03 | 1.93 | 2.26 | 1.24 | 1.57 | 1.43 | 1.35 | 1.40 | 1.33 | 1.42 | 2.38 | 2.88 | 2.46 | 1.70 | 1.49 | 1.27 | 1.83 |
| | MS | 1.88 | 1.84 | 2.05 | 1.86 | 1.85 | 1.90 | 1.87 | 1.11 | 1.43 | 1.04 | 1.05 | 1.25 | 0.98 | 1.09 | 1.71 | 1.55 | 1.39 | 1.38 | 2.12 | 1.99 | 2.74 |
| | MV | 1.30 | 1.32 | 1.10 | 1.13 | 1.24 | 1.13 | 1.11 | 1.31 | 1.70 | 1.46 | 1.43 | 1.52 | 1.36 | 1.50 | 2.32 | 2.43 | 2.49 | 2.10 | 1.62 | 2.24 | 2.21 |
| | SV | 1.43 | 1.39 | 1.01 | 1.23 | 1.35 | 1.23 | 1.15 | 2.08 | 1.81 | 2.31 | 2.11 | 1.90 | 2.26 | 2.11 | 2.04 | 2.45 | 3.01 | 2.64 | 2.23 | 0.61 | 1.04 |

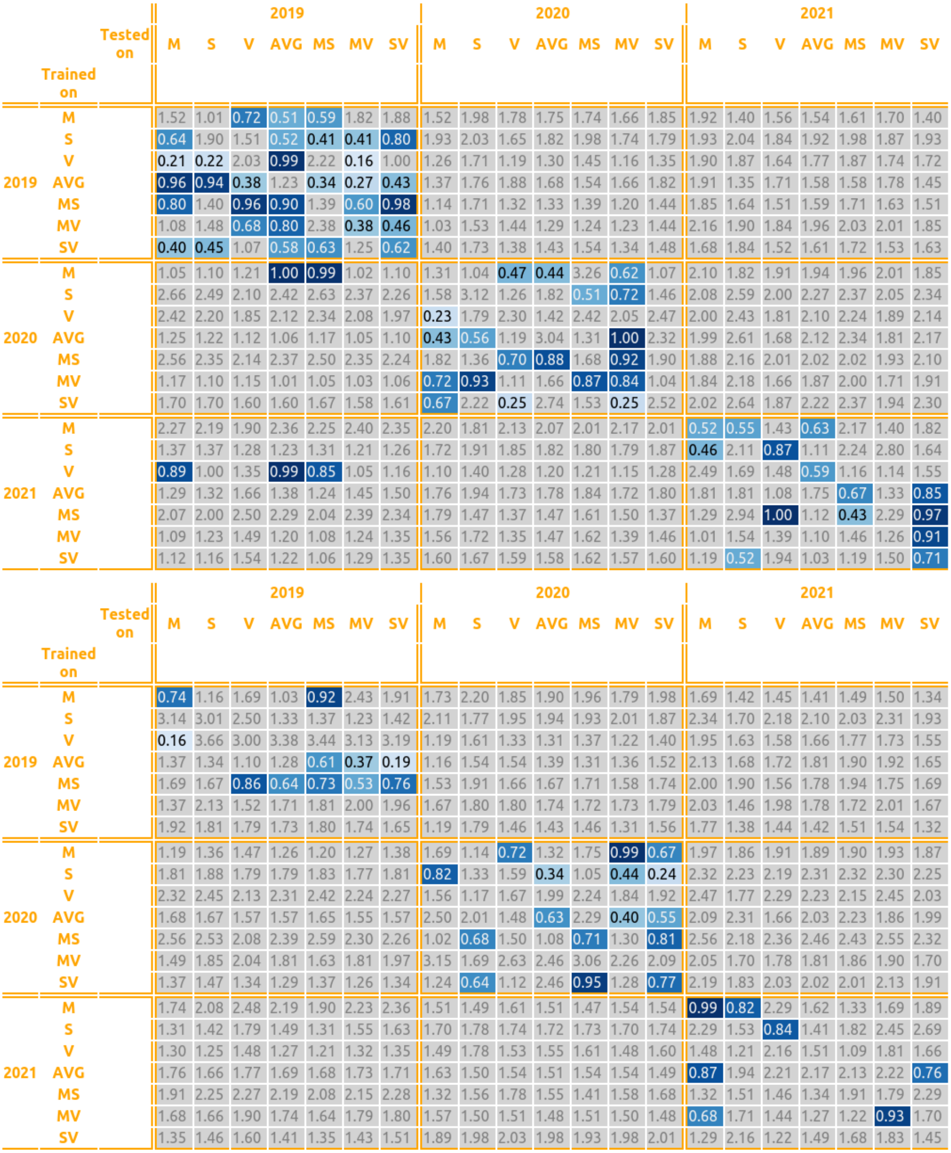
Figure A.2: Performance (SSE residual $r^2$) of the variety-specific Challenger model (top) and the variety-agnostic model tested on the Challenger variety (bottom). Rows correspond to the training dataset, columns show the testing configuration. SSE residuals $r^2 \geq 1$ are grayed out.



| Trained on | Tested on | 2019 | | | | | | | 2020 | | | | | | | 2021 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | M | S | V | AVG | MS | MV | SV | M | S | V | AVG | MS | MV | SV | M | S | V | AVG | MS | MV | SV |
| 2019 | M | | | | | | 0.90 | 0.73 | | | | | | | | | | | | | | |
| | S | | | | | | | | | | | | | | | | | | | | | |
| | V | | 0.71 | | | | | | | | | | | | | | | | | | | |
| | AVG | | 0.94 | | | 0.82 | 0.72 | 0.89 | | | | | | | | 0.40 | | | | 0.42 | | |
| | MS | | | | | 0.75 | 0.71 | 0.94 | | | | | | | | 0.39 | | | | | | |
| | MV | | | | | 0.77 | 0.89 | 0.93 | | | | | | | | 0.37 | | | | | | |
| | SV | | | | | | | 0.93 | | | | | | | | | | | | | | |
| 2020 | M | | | | | | | | | | | | 0.90 | | 0.77 | | | | | | | |
| | S | | | | | | | | | | | | | | 0.76 | | | | | | | |
| | V | 0.37 | | | 0.39 | 0.39 | 0.37 | | | | 0.87 | | | | 0.71 | | | | | | | |
| | AVG | | | -0.44 | | | | | 0.73 | 0.76 | | | 0.91 | | | | | | | | | |
| | MS | | | | | | | | | | | | | | | | | | | | | |
| | MV | | | | | | | | 0.92 | | | 0.90 | 0.78 | 0.70 | | | | | | | | |
| | SV | | | | | | | | | 0.87 | 0.73 | | 0.79 | | | | | | | | | |
| 2021 | M | 0.50 | | | 0.43 | 0.48 | 0.42 | | | | | | | | | | | | | | | |
| | S | | | | | | | | | | -0.39 | | | | | | 0.76 | | | | | -0.61 |
| | V | | | | | | | | | | | | | | | | | | | | | |
| | AVG | 0.42 | 0.49 | | 0.48 | 0.49 | 0.40 | 0.46 | | | | | | | | | | | | | | |
| | MS | 0.37 | | | 0.37 | | 0.39 | | | | -0.45 | -0.39 | | -0.43 | -0.41 | | | | | 0.97 | | |
| | MV | | | | | | | | | | | | | | | | | | | | | |
| | SV | | | | | | | | | | | | | | | | 0.81 | | | | | |

| Trained on | Tested on | 2019 | | | | | | | 2020 | | | | | | | 2021 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | M | S | V | AVG | MS | MV | SV | M | S | V | AVG | MS | MV | SV | M | S | V | AVG | MS | MV | SV |
| 2019 | M | | | | | | | | | | -0.42 | | | | | | | | 0.47 | 0.43 | 0.38 | |
| | S | | -0.74 | | | | | | 0.37 | 0.39 | 0.44 | 0.48 | 0.42 | 0.46 | 0.48 | | | | | | | |
| | V | | | 0.72 | | | | | | | | | | | | | | | | | | |
| | AVG | | | | | | | | | 0.45 | | 0.39 | 0.44 | | 0.38 | | | | | | | |
| | MS | | | | | | | | | | | | | | | 0.39 | | | | | | |
| | MV | | | | | | | | 0.49 | 0.58 | | 0.44 | 0.60 | | 0.40 | | | | | | | |
| | SV | | | | | | | | | | | | | | | 0.44 | | | 0.40 | 0.41 | 0.44 | |
| 2020 | M | | | 0.40 | 0.38 | | 0.39 | 0.37 | | | | | | | | | | | | | | |
| | S | | | 0.45 | | | | | | | | | | | | | -0.37 | | | | | -0.40 |
| | V | | | | | | | 0.38 | | | | | | | | | | | | | | |
| | AVG | | | 0.50 | 0.39 | | 0.43 | 0.39 | | | | | | | | | -0.40 | | | | | -0.44 |
| | MS | | | 0.53 | | | | | | | | | 0.73 | | | | -0.40 | | | | | |
| | MV | 0.36 | 0.42 | 0.47 | 0.48 | 0.43 | 0.45 | 0.52 | | | | | | | | | | | | | | |
| | SV | | | | 0.38 | | | 0.41 | | | | | | | | | -0.43 | | | | | -0.48 |
| 2021 | M | | | | | | | | | | 0.44 | 0.39 | | 0.43 | 0.39 | | | | | | | |
| | S | | | | | | | | 0.50 | 0.42 | | | 0.49 | | | | | | | | | |
| | V | | | | | | | | | | | | | | | | | | | | | |
| | AVG | | | | | | | | 0.38 | | | | | | | | | | | | | |
| | MS | | | | | | | | 0.45 | | 0.48 | 0.47 | 0.37 | 0.51 | 0.45 | | | | | | | |
| | MV | | | 0.45 | 0.44 | 0.38 | 0.43 | 0.44 | | | | | | | | | | | | | | |
| | SV | | | 0.50 | 0.39 | | 0.38 | 0.43 | | | | | | | | | | | | | 0.69 | |

Figure A.3: Performance (Pearson correlation $c$) of the variety-specific Challenger model (top) and the variety-agnostic model tested on the Challenger variety (bottom). Rows correspond to the training dataset, columns show the testing configuration. Correlations with $p \geq 0.05$ are omitted.



| Trained on | Tested on | 2019 | | | | | | | 2020 | | | | | | | 2021 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | M | S | V | AVG | MS | MV | SV | M | S | V | AVG | MS | MV | SV | M | S | V | AVG | MS | MV | SV |
| 2019 | M | 1.52 | 1.01 | 0.72 | 0.51 | 0.59 | 1.82 | 1.88 | 1.52 | 1.98 | 1.78 | 1.75 | 1.74 | 1.66 | 1.85 | 1.92 | 1.40 | 1.56 | 1.54 | 1.61 | 1.70 | 1.40 |
| | S | 0.64 | 1.90 | 1.51 | 0.52 | 0.41 | 0.41 | 0.80 | 1.93 | 2.03 | 1.65 | 1.82 | 1.98 | 1.74 | 1.79 | 1.93 | 2.04 | 1.84 | 1.92 | 1.98 | 1.87 | 1.93 |
| | V | 0.21 | 0.22 | 2.03 | 0.99 | 2.22 | 0.16 | 1.00 | 1.26 | 1.71 | 1.19 | 1.30 | 1.45 | 1.16 | 1.35 | 1.90 | 1.87 | 1.64 | 1.77 | 1.87 | 1.74 | 1.72 |
| | AVG | 0.96 | 0.94 | 0.38 | 1.23 | 0.34 | 0.27 | 0.43 | 1.37 | 1.76 | 1.88 | 1.68 | 1.54 | 1.66 | 1.82 | 1.91 | 1.35 | 1.71 | 1.58 | 1.58 | 1.78 | 1.45 |
| | MS | 0.80 | 1.40 | 0.96 | 0.90 | 1.39 | 0.60 | 0.98 | 1.14 | 1.71 | 1.32 | 1.33 | 1.39 | 1.20 | 1.44 | 1.85 | 1.64 | 1.51 | 1.59 | 1.71 | 1.63 | 1.51 |
| | MV | 1.08 | 1.48 | 0.68 | 0.80 | 2.38 | 0.38 | 0.46 | 1.03 | 1.53 | 1.44 | 1.29 | 1.24 | 1.23 | 1.44 | 2.16 | 1.90 | 1.84 | 1.96 | 2.03 | 2.01 | 1.85 |
| | SV | 0.40 | 0.45 | 1.07 | 0.58 | 0.63 | 1.25 | 0.62 | 1.40 | 1.73 | 1.38 | 1.43 | 1.54 | 1.34 | 1.48 | 1.68 | 1.84 | 1.52 | 1.61 | 1.72 | 1.53 | 1.63 |
| 2020 | M | 1.05 | 1.10 | 1.21 | 1.00 | 0.99 | 1.02 | 1.10 | 1.31 | 1.04 | 0.47 | 0.44 | 3.26 | 0.62 | 1.07 | 2.10 | 1.82 | 1.91 | 1.94 | 1.96 | 2.01 | 1.85 |
| | S | 2.66 | 2.49 | 2.10 | 2.42 | 2.63 | 2.37 | 2.26 | 1.58 | 3.12 | 1.26 | 1.82 | 0.51 | 0.72 | 1.46 | 2.08 | 2.59 | 2.00 | 2.27 | 2.37 | 2.05 | 2.34 |
| | V | 2.42 | 2.20 | 1.85 | 2.12 | 2.34 | 2.08 | 1.97 | 0.23 | 1.79 | 2.30 | 1.42 | 2.42 | 2.05 | 2.47 | 2.00 | 2.43 | 1.81 | 2.10 | 2.24 | 1.89 | 2.14 |
| | AVG | 1.25 | 1.22 | 1.12 | 1.06 | 1.17 | 1.05 | 1.10 | 0.43 | 0.56 | 1.19 | 3.04 | 1.31 | 1.00 | 2.32 | 1.99 | 2.61 | 1.68 | 2.12 | 2.34 | 1.81 | 2.17 |
| | MS | 2.56 | 2.35 | 2.14 | 2.37 | 2.50 | 2.35 | 2.24 | 1.82 | 1.36 | 0.70 | 0.88 | 1.68 | 0.92 | 1.90 | 1.88 | 2.16 | 2.01 | 2.02 | 2.02 | 1.93 | 2.10 |
| | MV | 1.17 | 1.10 | 1.15 | 1.01 | 1.05 | 1.03 | 1.06 | 0.72 | 0.93 | 1.11 | 1.66 | 0.87 | 0.84 | 1.04 | 1.84 | 2.18 | 1.66 | 1.87 | 2.00 | 1.71 | 1.91 |
| | SV | 1.70 | 1.70 | 1.60 | 1.60 | 1.67 | 1.58 | 1.61 | 0.67 | 2.22 | 0.25 | 2.74 | 1.53 | 0.25 | 2.52 | 2.02 | 2.64 | 1.87 | 2.22 | 2.37 | 1.94 | 2.30 |
| 2021 | M | 2.27 | 2.19 | 1.90 | 2.36 | 2.25 | 2.40 | 2.35 | 2.20 | 1.81 | 2.13 | 2.07 | 2.01 | 2.17 | 2.01 | 0.52 | 0.55 | 1.43 | 0.63 | 2.17 | 1.40 | 1.82 |
| | S | 1.37 | 1.37 | 1.28 | 1.23 | 1.31 | 1.21 | 1.26 | 1.72 | 1.91 | 1.85 | 1.82 | 1.80 | 1.79 | 1.87 | 0.46 | 2.11 | 0.87 | 1.11 | 2.24 | 2.80 | 1.64 |
| | V | 0.89 | 1.00 | 1.35 | 0.99 | 0.85 | 1.05 | 1.16 | 1.10 | 1.40 | 1.28 | 1.20 | 1.21 | 1.15 | 1.28 | 2.49 | 1.69 | 1.48 | 0.59 | 1.16 | 1.14 | 1.55 |
| | AVG | 1.29 | 1.32 | 1.66 | 1.38 | 1.24 | 1.45 | 1.50 | 1.76 | 1.94 | 1.73 | 1.78 | 1.84 | 1.72 | 1.80 | 1.81 | 1.81 | 1.08 | 1.75 | 0.67 | 1.33 | 0.85 |
| | MS | 2.07 | 2.00 | 2.50 | 2.29 | 2.04 | 2.39 | 2.34 | 1.79 | 1.47 | 1.37 | 1.47 | 1.61 | 1.50 | 1.37 | 1.29 | 2.94 | 1.00 | 1.12 | 0.43 | 2.29 | 0.97 |
| | MV | 1.09 | 1.23 | 1.49 | 1.20 | 1.08 | 1.24 | 1.35 | 1.56 | 1.72 | 1.35 | 1.47 | 1.62 | 1.39 | 1.46 | 1.01 | 1.54 | 1.39 | 1.10 | 1.46 | 1.26 | 0.91 |
| | SV | 1.12 | 1.16 | 1.54 | 1.22 | 1.06 | 1.29 | 1.35 | 1.60 | 1.67 | 1.59 | 1.58 | 1.62 | 1.57 | 1.60 | 1.19 | 0.52 | 1.94 | 1.03 | 1.19 | 1.50 | 0.71 |

| Trained on | Tested on | 2019 | | | | | | | 2020 | | | | | | | 2021 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | M | S | V | AVG | MS | MV | SV | M | S | V | AVG | MS | MV | SV | M | S | V | AVG | MS | MV | SV |
| 2019 | M | 0.74 | 1.16 | 1.69 | 1.03 | 0.92 | 2.43 | 1.91 | 1.73 | 2.20 | 1.85 | 1.90 | 1.96 | 1.79 | 1.98 | 1.69 | 1.42 | 1.45 | 1.41 | 1.49 | 1.50 | 1.34 |
| | S | 3.14 | 3.01 | 2.50 | 1.33 | 1.37 | 1.23 | 1.42 | 2.11 | 1.77 | 1.95 | 1.94 | 1.93 | 2.01 | 1.87 | 2.34 | 1.70 | 2.18 | 2.10 | 2.03 | 2.31 | 1.93 |
| | V | 0.16 | 3.66 | 3.00 | 3.38 | 3.44 | 3.13 | 3.19 | 1.19 | 1.61 | 1.33 | 1.31 | 1.37 | 1.22 | 1.40 | 1.95 | 1.63 | 1.58 | 1.66 | 1.77 | 1.73 | 1.55 |
| | AVG | 1.37 | 1.34 | 1.10 | 1.28 | 0.61 | 0.37 | 0.19 | 1.16 | 1.54 | 1.54 | 1.39 | 1.31 | 1.36 | 1.52 | 2.13 | 1.68 | 1.72 | 1.81 | 1.90 | 1.92 | 1.65 |
| | MS | 1.69 | 1.67 | 0.86 | 0.64 | 0.73 | 0.53 | 0.76 | 1.53 | 1.91 | 1.66 | 1.67 | 1.71 | 1.58 | 1.74 | 2.00 | 1.90 | 1.56 | 1.78 | 1.94 | 1.75 | 1.69 |
| | MV | 1.37 | 2.13 | 1.52 | 1.71 | 1.81 | 2.00 | 1.96 | 1.67 | 1.80 | 1.80 | 1.74 | 1.72 | 1.73 | 1.79 | 2.03 | 1.46 | 1.98 | 1.78 | 1.72 | 2.01 | 1.67 |
| | SV | 1.92 | 1.81 | 1.79 | 1.73 | 1.80 | 1.74 | 1.65 | 1.19 | 1.79 | 1.46 | 1.43 | 1.46 | 1.31 | 1.56 | 1.77 | 1.38 | 1.44 | 1.42 | 1.51 | 1.54 | 1.32 |
| 2020 | M | 1.19 | 1.36 | 1.47 | 1.26 | 1.20 | 1.27 | 1.38 | 1.69 | 1.14 | 0.72 | 1.32 | 1.75 | 0.99 | 0.67 | 1.97 | 1.86 | 1.91 | 1.89 | 1.90 | 1.93 | 1.87 |
| | S | 1.81 | 1.88 | 1.79 | 1.79 | 1.83 | 1.77 | 1.81 | 0.82 | 1.33 | 1.59 | 0.34 | 1.05 | 0.44 | 0.24 | 2.32 | 2.23 | 2.19 | 2.31 | 2.32 | 2.30 | 2.25 |
| | V | 2.32 | 2.45 | 2.13 | 2.31 | 2.42 | 2.24 | 2.27 | 1.56 | 1.17 | 1.67 | 1.99 | 2.24 | 1.84 | 1.92 | 2.47 | 1.77 | 2.29 | 2.23 | 2.15 | 2.45 | 2.03 |
| | AVG | 1.68 | 1.67 | 1.57 | 1.57 | 1.65 | 1.55 | 1.57 | 2.50 | 2.01 | 1.48 | 0.63 | 2.29 | 0.40 | 0.55 | 2.09 | 2.31 | 1.66 | 2.03 | 2.23 | 1.86 | 1.99 |
| | MS | 2.56 | 2.53 | 2.08 | 2.39 | 2.59 | 2.30 | 2.26 | 1.02 | 0.68 | 1.50 | 1.08 | 0.71 | 1.30 | 0.81 | 2.56 | 2.18 | 2.36 | 2.46 | 2.43 | 2.55 | 2.32 |
| | MV | 1.49 | 1.85 | 2.04 | 1.81 | 1.63 | 1.81 | 1.97 | 3.15 | 1.69 | 2.63 | 2.46 | 3.06 | 2.26 | 2.09 | 2.05 | 1.70 | 1.78 | 1.81 | 1.86 | 1.90 | 1.70 |
| | SV | 1.37 | 1.47 | 1.34 | 1.29 | 1.37 | 1.26 | 1.34 | 1.24 | 0.64 | 1.12 | 2.46 | 0.95 | 1.28 | 0.77 | 2.19 | 1.83 | 2.03 | 2.02 | 2.01 | 2.13 | 1.91 |
| 2021 | M | 1.74 | 2.08 | 2.48 | 2.19 | 1.90 | 2.23 | 2.36 | 1.51 | 1.49 | 1.61 | 1.51 | 1.47 | 1.54 | 1.54 | 0.99 | 0.82 | 2.29 | 1.62 | 1.33 | 1.69 | 1.89 |
| | S | 1.31 | 1.42 | 1.79 | 1.49 | 1.31 | 1.55 | 1.63 | 1.70 | 1.78 | 1.74 | 1.72 | 1.73 | 1.70 | 1.74 | 2.29 | 1.53 | 0.84 | 1.41 | 1.82 | 2.45 | 2.69 |
| | V | 1.30 | 1.25 | 1.48 | 1.27 | 1.21 | 1.32 | 1.35 | 1.49 | 1.78 | 1.53 | 1.55 | 1.61 | 1.48 | 1.60 | 1.48 | 1.21 | 2.16 | 1.51 | 1.09 | 1.81 | 1.66 |
| | AVG | 1.76 | 1.66 | 1.77 | 1.69 | 1.68 | 1.73 | 1.71 | 1.63 | 1.50 | 1.54 | 1.51 | 1.54 | 1.54 | 1.49 | 0.87 | 1.94 | 2.21 | 2.17 | 2.13 | 2.22 | 0.76 |
| | MS | 1.91 | 2.25 | 2.27 | 2.19 | 2.08 | 2.15 | 2.28 | 1.32 | 1.56 | 1.78 | 1.55 | 1.41 | 1.58 | 1.68 | 1.32 | 1.51 | 1.46 | 1.34 | 1.91 | 1.79 | 2.29 |
| | MV | 1.68 | 1.66 | 1.90 | 1.74 | 1.64 | 1.79 | 1.80 | 1.57 | 1.50 | 1.51 | 1.48 | 1.51 | 1.50 | 1.48 | 0.68 | 1.71 | 1.44 | 1.27 | 1.22 | 0.93 | 1.70 |
| | SV | 1.35 | 1.46 | 1.60 | 1.41 | 1.35 | 1.43 | 1.51 | 1.89 | 1.98 | 2.03 | 1.98 | 1.93 | 1.98 | 2.01 | 1.29 | 2.16 | 1.22 | 1.49 | 1.68 | 1.83 | 1.45 |

Figure A.4: Performance (SSE residual $r^2$) of the variety-specific Colomba model (top) and the variety-agnostic model tested on the Colomba variety (bottom). Rows correspond to the training dataset, columns show the testing configuration. SSE residuals $r^2 \geq 1$ are grayed out.

| | Tested on | 2019 | | | | | | | 2020 | | | | | | | 2021 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Trained on | | M | S | V | AVG | MS | MV | SV | M | S | V | AVG | MS | MV | SV | M | S | V | AVG | MS | MV | SV |
| 2019 | M | | | | 0.75 | 0.70 | | | | | | | | | | | | | | | | |
| | S | 0.68 | | | 0.74 | 0.80 | 0.80 | | | | | | | | | | | | | | | |
| | V | 0.90 | 0.89 | | | | 0.92 | | 0.37 | | 0.40 | | | 0.42 | | | | | | | | |
| | AVG | | | 0.81 | | 0.83 | 0.87 | 0.79 | | | | | | | | | | | | | | |
| | MS | | | | | | | | 0.43 | | | | | 0.40 | | | | | | | | |
| | MV | | | | | | 0.81 | 0.77 | 0.48 | | | | 0.38 | 0.38 | | | | | | | | |
| | SV | 0.80 | 0.78 | | 0.71 | 0.68 | | 0.69 | | | | | | | | | | | | | | |
| 2020 | M | 0.47 | 0.45 | 0.40 | 0.50 | 0.50 | 0.49 | 0.45 | | | 0.77 | 0.78 | | | | | | | | | | |
| | S | | | | | | | | | | | | 0.74 | | | | | | | | | |
| | V | | | | | | | | 0.88 | | | | | | | | | | | | | |
| | AVG | 0.38 | 0.39 | 0.44 | 0.47 | 0.42 | 0.48 | 0.45 | 0.79 | 0.72 | | | | | | | | | | | | |
| | MS | | | | | | | | | | | | | | | | | | | | | |
| | MV | 0.42 | 0.45 | 0.43 | 0.50 | 0.47 | 0.48 | 0.47 | | | | | | | | | | | | | | |
| | SV | | | | | | | | | | 0.88 | | | 0.88 | | | | | | | | |
| 2021 | M | | | | | | | | | | | | | | | | 0.72 | | 0.69 | | | |
| | S | | | | 0.39 | | 0.39 | 0.37 | | | | | | | | 0.77 | | | | | | |
| | V | 0.55 | 0.50 | | 0.50 | 0.57 | 0.48 | 0.42 | 0.45 | | 0.36 | 0.40 | 0.40 | 0.42 | | | | | 0.70 | | | |
| | AVG | | | | | 0.38 | | | | | | | | | | | | | | 0.67 | | |
| | MS | | | | | | | | | | | | | | | | | | | 0.79 | | |
| | MV | 0.46 | 0.38 | | 0.40 | 0.46 | 0.38 | | | | | | | | | | | | | | | |
| | SV | 0.44 | 0.42 | | 0.39 | 0.47 | | | | | | | | | | | 0.74 | | | | | 0.64 |

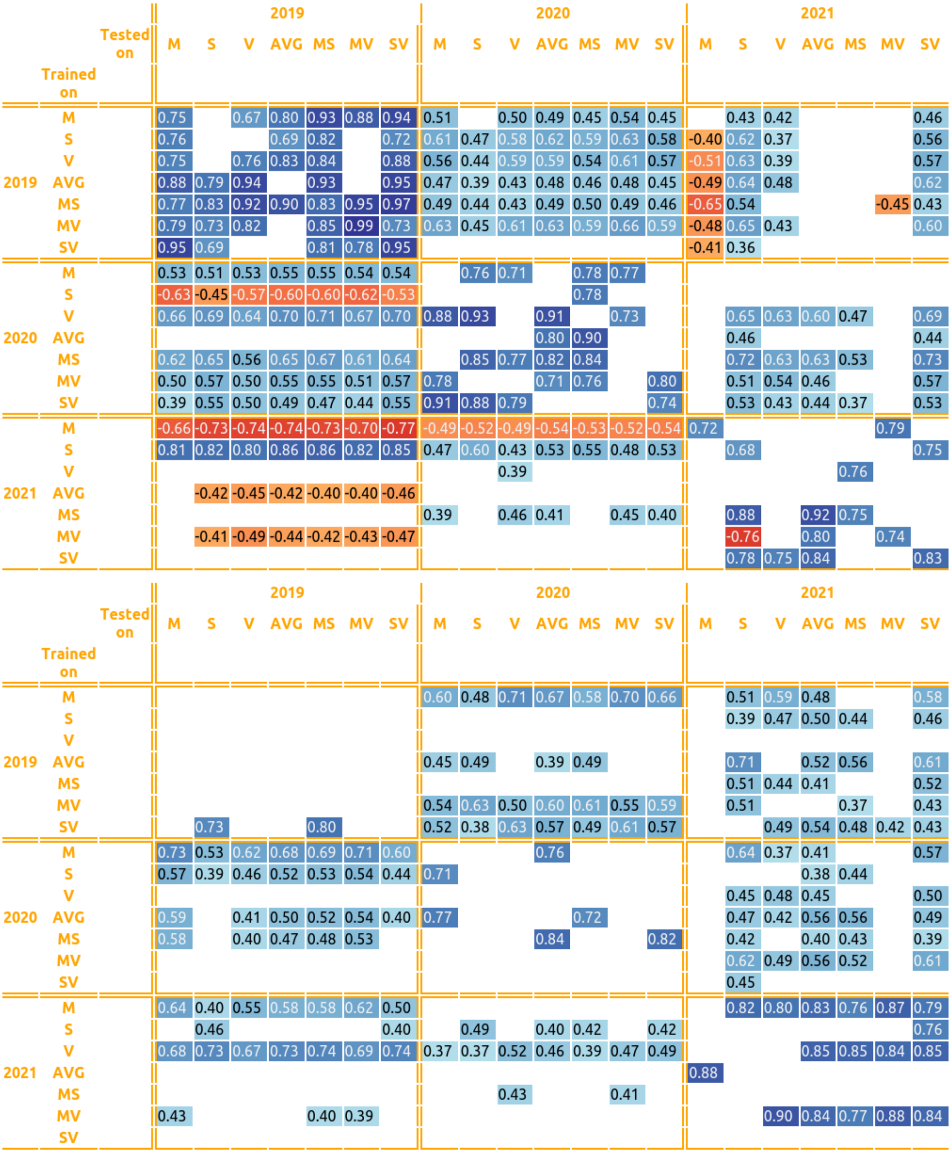
| | Tested on | 2019 | | | | | | | 2020 | | | | | | | 2021 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Trained on | | M | S | V | AVG | MS | MV | SV | M | S | V | AVG | MS | MV | SV | M | S | V | AVG | MS | MV | SV |
| 2019 | M | | | | | | | | | | | | | | | | | | | | | |
| | S | | | | | | | | | | | | | | | | | | | | | |
| | V | 0.92 | -0.83 | | | -0.72 | | | 0.41 | | | | | 0.39 | | | | | | | | |
| | AVG | | | | | | 0.81 | 0.90 | 0.42 | | | | | | | | | | | | | |
| | MS | | | | | | 0.74 | | | | | | | | | | | | | | | |
| | MV | | | | | | | | | | | | | | | | | | | | | |
| | SV | | | | | | | | 0.41 | | | | | | | | | | | | | |
| 2020 | M | 0.41 | | | 0.37 | 0.40 | | | | | | | | | | | | | | | | |
| | S | | | | | | | | | | | 0.83 | | 0.78 | 0.88 | | | | | | | |
| | V | | | | | | | | | | | | | | | | | | | | | |
| | AVG | | | | | | | | | | | | | 0.80 | 0.73 | | | | | | | |
| | MS | | | | | | | | | | | | | | | | | | | | | |
| | MV | | | | | | | | | | | | | | | | | | | | | |
| | SV | | | | | | 0.37 | | | | | | | | | | | | | | | |
| 2021 | M | | | | | | | | | | | | | | | | | | | | | |
| | S | | | | | | | | | | | | | | | | | | | | | |
| | V | | 0.38 | | | 0.40 | | | | | | | | | | | | | | | | |
| | AVG | | | | | | | | | | | | | | | | | | | | | |
| | MS | | | | | | | | | | | | | | | | | | | | | |
| | MV | | | | | | | | | | | | | | | | | | | | | |
| | SV | | | | | | | | | | | | | | | | | | | | | |

Figure A.5: Performance (Pearson correlation $c$) of the variety-specific Colomba model (top) and the variety-agnostic model tested on the Colomba variety (bottom). Rows correspond to the training dataset, columns show the testing configuration. Correlations with $p \geq 0.05$ are omitted.

| Trained on | Tested on | 2019 | | | | | | | 2020 | | | | | | | 2021 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | M | S | V | AVG | MS | MV | SV | M | S | V | AVG | MS | MV | SV | M | S | V | AVG | MS | MV | SV |
| 2019 | M | 0.50 | 0.70 | 0.66 | 0.41 | 0.14 | 0.24 | 0.12 | 0.98 | 1.40 | 1.00 | 1.01 | 1.10 | 0.93 | 1.10 | 2.55 | 1.15 | 1.16 | 1.39 | 1.66 | 1.96 | 1.09 |
| | S | 0.48 | 1.86 | 0.87 | 0.62 | 0.37 | 1.02 | 0.56 | 0.78 | 1.06 | 0.84 | 0.77 | 0.83 | 0.73 | 0.84 | 2.80 | 0.75 | 1.26 | 1.33 | 1.50 | 2.24 | 0.88 |
| | V | 0.50 | 1.65 | 0.47 | 0.33 | 0.32 | 0.74 | 0.24 | 0.87 | 1.11 | 0.82 | 0.81 | 0.91 | 0.77 | 0.86 | 3.02 | 0.73 | 1.23 | 1.45 | 1.71 | 2.42 | 0.85 |
| | AVG | 0.25 | 0.41 | 0.13 | 1.37 | 0.14 | 0.67 | 0.10 | 1.06 | 1.22 | 1.13 | 1.05 | 1.07 | 1.04 | 1.11 | 2.98 | 0.73 | 1.03 | 1.33 | 1.66 | 2.26 | 0.76 |
| | MS | 0.46 | 0.35 | 0.15 | 0.20 | 0.34 | 0.10 | 0.06 | 1.02 | 1.12 | 1.15 | 1.01 | 1.01 | 1.02 | 1.07 | 3.31 | 0.91 | 1.62 | 1.94 | 2.19 | 2.91 | 1.14 |
| | MV | 0.41 | 0.53 | 0.36 | 1.87 | 0.30 | 0.01 | 0.55 | 0.74 | 1.10 | 0.78 | 0.74 | 0.82 | 0.68 | 0.83 | 2.95 | 0.70 | 1.15 | 1.34 | 1.60 | 2.31 | 0.80 |
| | SV | 0.10 | 0.62 | 2.07 | 0.75 | 0.38 | 0.45 | 0.10 | 2.18 | 2.17 | 2.21 | 2.21 | 2.18 | 2.21 | 2.21 | 2.81 | 1.28 | 1.80 | 1.95 | 2.07 | 2.59 | 1.45 |
| 2020 | M | 0.94 | 0.98 | 0.95 | 0.89 | 0.90 | 0.92 | 0.91 | 2.04 | 0.48 | 0.58 | 1.66 | 0.45 | 0.46 | 0.64 | 2.39 | 1.38 | 2.02 | 1.85 | 1.75 | 2.36 | 1.61 |
| | S | 3.27 | 2.91 | 3.14 | 3.20 | 3.19 | 3.25 | 3.07 | 0.99 | 2.10 | 1.04 | 1.40 | 0.44 | 1.59 | 1.32 | 2.15 | 1.34 | 1.81 | 1.56 | 1.46 | 2.01 | 1.50 |
| | V | 0.68 | 0.61 | 0.72 | 0.59 | 0.58 | 0.67 | 0.59 | 0.23 | 0.13 | 1.42 | 0.18 | 0.75 | 0.54 | 0.96 | 2.43 | 0.70 | 0.74 | 0.81 | 1.06 | 1.59 | 0.62 |
| | AVG | 2.57 | 2.12 | 2.33 | 2.41 | 2.42 | 2.50 | 2.23 | 0.82 | 1.28 | 2.17 | 0.40 | 0.19 | 1.09 | 0.69 | 2.47 | 1.08 | 1.34 | 1.37 | 1.51 | 2.00 | 1.12 |
| | MS | 0.76 | 0.69 | 0.88 | 0.70 | 0.67 | 0.78 | 0.72 | 1.60 | 0.30 | 0.46 | 0.36 | 0.32 | 1.21 | 0.73 | 2.46 | 0.56 | 0.75 | 0.73 | 0.94 | 1.62 | 0.53 |
| | MV | 1.01 | 0.86 | 1.00 | 0.90 | 0.89 | 0.98 | 0.87 | 0.44 | 0.80 | 0.75 | 0.58 | 0.49 | 1.28 | 0.39 | 2.45 | 0.97 | 0.92 | 1.09 | 1.37 | 1.72 | 0.87 |
| | SV | 1.22 | 0.91 | 1.00 | 1.02 | 1.05 | 1.13 | 0.90 | 0.18 | 0.24 | 0.43 | 1.35 | 0.89 | 0.89 | 0.52 | 2.36 | 0.94 | 1.15 | 1.12 | 1.25 | 1.79 | 0.95 |
| 2021 | M | 3.32 | 3.46 | 3.48 | 3.48 | 3.45 | 3.41 | 3.54 | 2.99 | 3.05 | 2.98 | 3.09 | 3.07 | 3.05 | 3.08 | 0.56 | 1.72 | 2.64 | 2.18 | 1.99 | 0.42 | 3.08 |
| | S | 0.39 | 0.37 | 0.40 | 0.29 | 0.29 | 0.35 | 0.30 | 1.06 | 0.80 | 1.14 | 0.94 | 0.90 | 1.04 | 0.94 | 3.09 | 0.64 | 1.38 | 2.19 | 1.34 | 1.83 | 0.50 |
| | V | 2.23 | 1.98 | 2.36 | 2.20 | 2.14 | 2.28 | 2.16 | 1.75 | 1.94 | 1.22 | 1.56 | 1.81 | 1.45 | 1.46 | 1.09 | 1.13 | 2.06 | 0.99 | 0.47 | 1.96 | 1.51 |
| | AVG | 2.71 | 2.85 | 2.90 | 2.84 | 2.80 | 2.79 | 2.92 | 1.67 | 1.66 | 1.48 | 1.56 | 1.65 | 1.55 | 1.51 | 2.70 | 0.88 | 1.06 | 1.41 | 0.74 | 1.53 | 0.96 |
| | MS | 2.54 | 2.64 | 2.77 | 2.66 | 2.61 | 2.63 | 2.74 | 1.23 | 1.50 | 1.08 | 1.17 | 1.30 | 1.10 | 1.19 | 1.34 | 0.25 | 1.94 | 0.16 | 0.50 | 1.46 | 0.87 |
| | MV | 2.77 | 2.83 | 2.98 | 2.89 | 2.84 | 2.86 | 2.94 | 2.01 | 2.21 | 1.64 | 1.91 | 2.10 | 1.81 | 1.85 | 0.80 | 3.52 | 1.47 | 0.41 | 0.96 | 0.51 | 2.04 |
| | SV | 1.34 | 2.10 | 1.78 | 1.65 | 1.61 | 1.48 | 1.95 | 1.47 | 1.73 | 1.74 | 1.61 | 1.55 | 1.58 | 1.72 | 1.89 | 0.43 | 0.50 | 0.32 | 0.73 | 2.82 | 0.34 |

| Trained on | Tested on | 2019 | | | | | | | 2020 | | | | | | | 2021 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | M | S | V | AVG | MS | MV | SV | M | S | V | AVG | MS | MV | SV | M | S | V | AVG | MS | MV | SV |
| 2019 | M | 1.66 | 2.06 | 1.05 | 1.36 | 1.78 | 2.31 | 2.66 | 0.80 | 1.04 | 0.59 | 0.67 | 0.84 | 0.61 | 0.67 | 2.41 | 0.99 | 0.82 | 1.03 | 1.35 | 1.63 | 0.84 |
| | S | 2.39 | 2.81 | 1.33 | 1.43 | 0.92 | 1.75 | 1.91 | 2.14 | 2.28 | 1.76 | 2.03 | 2.21 | 1.94 | 1.96 | 1.95 | 1.22 | 1.06 | 0.99 | 1.12 | 1.37 | 1.08 |
| | V | 0.70 | 3.15 | 1.65 | 2.22 | 2.70 | 1.88 | 1.93 | 1.62 | 1.60 | 1.46 | 1.51 | 1.59 | 1.51 | 1.48 | 1.61 | 1.76 | 1.51 | 1.35 | 1.35 | 1.35 | 1.63 |
| | AVG | 2.36 | 2.38 | 2.46 | 2.48 | 1.68 | 1.89 | 1.96 | 1.10 | 1.03 | 1.62 | 1.22 | 1.02 | 1.32 | 1.35 | 2.38 | 0.58 | 1.30 | 0.95 | 0.89 | 1.89 | 0.79 |
| | MS | 1.47 | 1.46 | 1.91 | 1.64 | 1.50 | 1.51 | 1.95 | 1.34 | 1.56 | 1.54 | 1.42 | 1.40 | 1.41 | 1.51 | 2.42 | 0.99 | 1.12 | 1.17 | 1.36 | 1.82 | 0.97 |
| | MV | 1.79 | 1.77 | 1.01 | 1.40 | 1.76 | 0.98 | 1.01 | 0.92 | 0.74 | 0.99 | 0.81 | 0.79 | 0.89 | 0.82 | 2.31 | 0.99 | 1.53 | 1.30 | 1.25 | 1.98 | 1.15 |
| | SV | 1.89 | 0.53 | 1.41 | 0.68 | 0.40 | 0.74 | 1.83 | 0.97 | 1.23 | 0.74 | 0.85 | 1.02 | 0.78 | 0.85 | 1.76 | 1.32 | 1.01 | 0.91 | 1.03 | 1.17 | 1.13 |
| 2020 | M | 0.54 | 0.94 | 0.76 | 0.63 | 0.62 | 0.59 | 0.80 | 1.34 | 1.64 | 1.83 | 0.48 | 1.76 | 1.84 | 1.74 | 2.61 | 0.73 | 1.26 | 1.19 | 1.28 | 2.07 | 0.86 |
| | S | 0.87 | 1.22 | 1.08 | 0.97 | 0.95 | 0.92 | 1.11 | 0.58 | 0.79 | 0.77 | 0.87 | 0.65 | 0.81 | 0.91 | 1.75 | 1.42 | 1.57 | 1.24 | 1.13 | 1.51 | 1.44 |
| | V | 1.93 | 2.49 | 2.25 | 2.18 | 2.15 | 2.04 | 2.40 | 1.12 | 1.21 | 2.09 | 1.66 | 1.18 | 1.79 | 1.84 | 2.25 | 1.09 | 1.04 | 1.10 | 1.30 | 1.62 | 1.00 |
| | AVG | 0.83 | 1.30 | 1.18 | 0.99 | 0.95 | 0.93 | 1.20 | 0.46 | 0.67 | 0.81 | 1.67 | 0.56 | 1.82 | 1.60 | 1.89 | 1.05 | 1.16 | 0.89 | 0.88 | 1.38 | 1.03 |
| | MS | 0.84 | 1.51 | 1.20 | 1.07 | 1.05 | 0.94 | 1.33 | 0.63 | 1.18 | 0.99 | 0.33 | 0.94 | 0.87 | 0.37 | 2.03 | 1.15 | 1.47 | 1.20 | 1.14 | 1.70 | 1.23 |
| | MV | 1.43 | 1.85 | 1.81 | 1.63 | 1.57 | 1.56 | 1.82 | 1.10 | 0.73 | 0.98 | 0.79 | 0.64 | 0.62 | 1.12 | 2.27 | 0.77 | 1.02 | 0.87 | 0.97 | 1.62 | 0.78 |
| | SV | 1.96 | 2.20 | 2.17 | 2.08 | 2.05 | 2.03 | 2.20 | 0.65 | 0.72 | 1.07 | 1.18 | 0.96 | 1.19 | 0.85 | 2.32 | 1.09 | 1.90 | 1.55 | 1.37 | 2.23 | 1.38 |
| 2021 | M | 0.72 | 1.19 | 0.90 | 0.84 | 0.84 | 0.76 | 1.01 | 1.44 | 1.55 | 1.53 | 1.46 | 1.45 | 1.45 | 1.51 | 1.74 | 0.36 | 0.41 | 0.34 | 0.49 | 0.27 | 0.42 |
| | S | 1.67 | 1.07 | 1.45 | 1.41 | 1.41 | 1.58 | 1.21 | 1.34 | 1.01 | 1.35 | 1.20 | 1.17 | 1.30 | 1.16 | 2.13 | 1.82 | 3.13 | 2.49 | 1.96 | 1.09 | 0.48 |
| | V | 0.64 | 0.54 | 0.66 | 0.54 | 0.53 | 0.62 | 0.52 | 1.26 | 1.26 | 0.96 | 1.07 | 1.22 | 1.05 | 1.01 | 1.36 | 1.08 | 0.74 | 0.29 | 0.30 | 0.33 | 0.30 |
| | AVG | 2.01 | 2.02 | 2.14 | 2.05 | 2.02 | 2.06 | 2.08 | 1.63 | 1.95 | 1.33 | 1.56 | 1.75 | 1.45 | 1.54 | 0.23 | 1.56 | 1.78 | 1.62 | 1.54 | 1.66 | 0.74 |
| | MS | 2.42 | 2.47 | 2.57 | 2.50 | 2.46 | 2.48 | 2.54 | 1.32 | 1.76 | 1.13 | 1.30 | 1.47 | 1.17 | 1.33 | 2.29 | 2.14 | 1.87 | 2.13 | 0.68 | 0.68 | 1.47 |
| | MV | 1.14 | 1.42 | 1.39 | 1.24 | 1.21 | 1.21 | 1.37 | 1.70 | 2.21 | 1.52 | 1.73 | 1.90 | 1.58 | 1.77 | 0.87 | 2.67 | 0.20 | 0.32 | 0.47 | 0.24 | 0.31 |
| | SV | 1.37 | 1.37 | 1.38 | 1.34 | 1.34 | 1.36 | 1.34 | 1.65 | 2.02 | 2.04 | 1.88 | 1.79 | 1.84 | 2.03 | 0.88 | 1.03 | 0.78 | 0.81 | 0.91 | 0.91 | 0.98 |

Figure A.6: Performance (SSE residual $r^2$) of the variety-specific Festien model (top) and the variety-agnostic model tested on the Festien variety (bottom). Rows correspond to the training dataset, columns show the testing configuration. SSE residuals $r^2 \geq 1$ are grayed out.



| | Tested on | 2019 | | | | | | | 2020 | | | | | | | 2021 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Trained on | | M | S | V | AVG | MS | MV | SV | M | S | V | AVG | MS | MV | SV | M | S | V | AVG | MS | MV | SV |
| 2019 | M | 0.75 | | 0.67 | 0.80 | 0.93 | 0.88 | 0.94 | 0.51 | | 0.50 | 0.49 | 0.45 | 0.54 | 0.45 | | 0.43 | 0.42 | | | | 0.46 |
| | S | 0.76 | | | 0.69 | 0.82 | | 0.72 | 0.61 | 0.47 | 0.58 | 0.62 | 0.59 | 0.63 | 0.58 | -0.40 | 0.62 | 0.37 | | | | 0.56 |
| | V | 0.75 | | 0.76 | 0.83 | 0.84 | | 0.88 | 0.56 | 0.44 | 0.59 | 0.59 | 0.54 | 0.61 | 0.57 | -0.51 | 0.63 | 0.39 | | | | 0.57 |
| | AVG | 0.88 | 0.79 | 0.94 | | 0.93 | | 0.95 | 0.47 | 0.39 | 0.43 | 0.48 | 0.46 | 0.48 | 0.45 | -0.49 | 0.64 | 0.48 | | | | 0.62 |
| | MS | 0.77 | 0.83 | 0.92 | 0.90 | 0.83 | 0.95 | 0.97 | 0.49 | 0.44 | 0.43 | 0.49 | 0.50 | 0.49 | 0.46 | -0.65 | 0.54 | | | | -0.45 | 0.43 |
| | MV | 0.79 | 0.73 | 0.82 | | 0.85 | 0.99 | 0.73 | 0.63 | 0.45 | 0.61 | 0.63 | 0.59 | 0.66 | 0.59 | -0.48 | 0.65 | 0.43 | | | | 0.60 |
| | SV | 0.95 | 0.69 | | | 0.81 | 0.78 | 0.95 | | | | | | | | -0.41 | 0.36 | | | | | |
| 2020 | M | 0.53 | 0.51 | 0.53 | 0.55 | 0.55 | 0.54 | 0.54 | | 0.76 | 0.71 | | 0.78 | 0.77 | | | | | | | | |
| | S | -0.63 | -0.45 | -0.57 | -0.60 | -0.60 | -0.62 | -0.53 | | | | | 0.78 | | | | | | | | | |
| | V | 0.66 | 0.69 | 0.64 | 0.70 | 0.71 | 0.67 | 0.70 | 0.88 | 0.93 | | 0.91 | | 0.73 | | | 0.65 | 0.63 | 0.60 | 0.47 | | 0.69 |
| | AVG | | | | | | | | | | | 0.80 | 0.90 | | | | 0.46 | | | | | 0.44 |
| | MS | 0.62 | 0.65 | 0.56 | 0.65 | 0.67 | 0.61 | 0.64 | | 0.85 | 0.77 | 0.82 | 0.84 | | | | 0.72 | 0.63 | 0.63 | 0.53 | | 0.73 |
| | MV | 0.50 | 0.57 | 0.50 | 0.55 | 0.55 | 0.51 | 0.57 | 0.78 | | | 0.71 | 0.76 | | 0.80 | | 0.51 | 0.54 | 0.46 | | | 0.57 |
| | SV | 0.39 | 0.55 | 0.50 | 0.49 | 0.47 | 0.44 | 0.55 | 0.91 | 0.88 | 0.79 | | | | 0.74 | | 0.53 | 0.43 | 0.44 | 0.37 | | 0.53 |
| 2021 | M | -0.66 | -0.73 | -0.74 | -0.74 | -0.73 | -0.70 | -0.77 | -0.49 | -0.52 | -0.49 | -0.54 | -0.53 | -0.52 | -0.54 | 0.72 | | | | | 0.79 | |
| | S | 0.81 | 0.82 | 0.80 | 0.86 | 0.86 | 0.82 | 0.85 | 0.47 | 0.60 | 0.43 | 0.53 | 0.55 | 0.48 | 0.53 | | 0.68 | | | | | 0.75 |
| | V | | | | | | | | | | 0.39 | | | | | | | | | 0.76 | | |
| | AVG | | -0.42 | -0.45 | -0.42 | -0.40 | -0.40 | -0.46 | | | | | | | | | | | | | | |
| | MS | | | | | | | | 0.39 | | 0.46 | 0.41 | | 0.45 | 0.40 | | 0.88 | | 0.92 | 0.75 | | |
| | MV | | -0.41 | -0.49 | -0.44 | -0.42 | -0.43 | -0.47 | | | | | | | | | -0.76 | | 0.80 | | 0.74 | |
| | SV | | | | | | | | | | | | | | | | 0.78 | 0.75 | 0.84 | | | 0.83 |

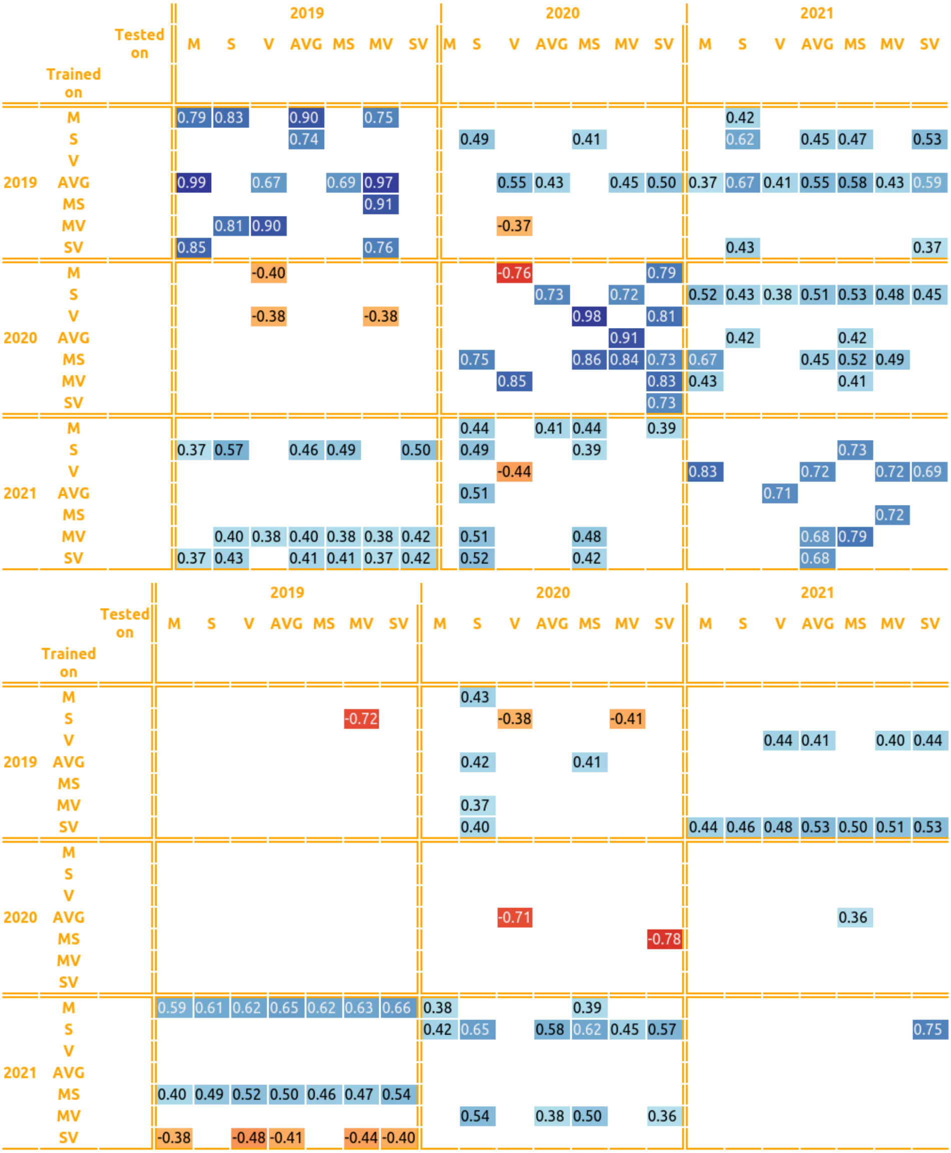
| | Tested on | 2019 | | | | | | | 2020 | | | | | | | 2021 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Trained on | | M | S | V | AVG | MS | MV | SV | M | S | V | AVG | MS | MV | SV | M | S | V | AVG | MS | MV | SV |
| 2019 | M | | | | | | | | 0.60 | 0.48 | 0.71 | 0.67 | 0.58 | 0.70 | 0.66 | | 0.51 | 0.59 | 0.48 | | | 0.58 |
| | S | | | | | | | | | | | | | | | | 0.39 | 0.47 | 0.50 | 0.44 | | 0.46 |
| | V | | | | | | | | | | | | | | | | | | | | | |
| | AVG | | | | | | | | 0.45 | 0.49 | | 0.39 | 0.49 | | | | 0.71 | | 0.52 | 0.56 | | 0.61 |
| | MS | | | | | | | | | | | | | | | | 0.51 | 0.44 | 0.41 | | | 0.52 |
| | MV | | | | | | | | 0.54 | 0.63 | 0.50 | 0.60 | 0.61 | 0.55 | 0.59 | | 0.51 | | | 0.37 | | 0.43 |
| | SV | | 0.73 | | | 0.80 | | | 0.52 | 0.38 | 0.63 | 0.57 | 0.49 | 0.61 | 0.57 | | | 0.49 | 0.54 | 0.48 | 0.42 | 0.43 |
| 2020 | M | 0.73 | 0.53 | 0.62 | 0.68 | 0.69 | 0.71 | 0.60 | | | | 0.76 | | | | | 0.64 | 0.37 | 0.41 | | | 0.57 |
| | S | 0.57 | 0.39 | 0.46 | 0.52 | 0.53 | 0.54 | 0.44 | 0.71 | | | | | | | | | | 0.38 | 0.44 | | |
| | V | | | | | | | | | | | | | | | | 0.45 | 0.48 | 0.45 | | | 0.50 |
| | AVG | 0.59 | | 0.41 | 0.50 | 0.52 | 0.54 | 0.40 | 0.77 | | | | 0.72 | | | | 0.47 | 0.42 | 0.56 | 0.56 | | 0.49 |
| | MS | 0.58 | | 0.40 | 0.47 | 0.48 | 0.53 | | | | | 0.84 | | | 0.82 | | 0.42 | | 0.40 | 0.43 | | 0.39 |
| | MV | | | | | | | | | | | | | | | | 0.62 | 0.49 | 0.56 | 0.52 | | 0.61 |
| | SV | | | | | | | | | | | | | | | | 0.45 | | | | | |
| 2021 | M | 0.64 | 0.40 | 0.55 | 0.58 | 0.58 | 0.62 | 0.50 | | | | | | | | | 0.82 | 0.80 | 0.83 | 0.76 | 0.87 | 0.79 |
| | S | | 0.46 | | | | | 0.40 | | 0.49 | | 0.40 | 0.42 | | 0.42 | | | | | | | 0.76 |
| | V | 0.68 | 0.73 | 0.67 | 0.73 | 0.74 | 0.69 | 0.74 | 0.37 | 0.37 | 0.52 | 0.46 | 0.39 | 0.47 | 0.49 | | | | 0.85 | 0.85 | 0.84 | 0.85 |
| | AVG | | | | | | | | | | | | | | | 0.88 | | | | | | |
| | MS | | | | | | | | | | 0.43 | | | 0.41 | | | | | | | | |
| | MV | 0.43 | | | | 0.40 | 0.39 | | | | | | | | | | | 0.90 | 0.84 | 0.77 | 0.88 | 0.84 |
| | SV | | | | | | | | | | | | | | | | | | | | | |

Figure A.7: Performance (Pearson correlation $c$) of the variety-specific Festien model (top) and the variety-agnostic model tested on the Festien variety (bottom). Rows correspond to the training dataset, columns show the testing configuration. Correlations with $p \geq 0.05$ are omitted.

| Trained on | Tested on | 2019 | | | | | | | 2020 | | | | | | | 2021 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | M | S | V | AVG | MS | MV | SV | M | S | V | AVG | MS | MV | SV | M | S | V | AVG | MS | MV | SV |
| 2019 | M | 0.42 | 0.34 | 1.54 | 0.20 | 0.85 | 0.51 | 1.42 | 2.08 | 2.11 | 2.30 | 2.22 | 2.11 | 2.27 | 2.24 | 1.70 | 1.16 | 1.54 | 1.39 | 1.36 | 1.57 | 1.29 |
| | S | 2.15 | 2.86 | 1.92 | 0.52 | 0.85 | 1.49 | 1.16 | 1.63 | 1.03 | 1.75 | 1.35 | 1.18 | 1.66 | 1.33 | 1.58 | 0.75 | 1.30 | 1.09 | 1.07 | 1.37 | 0.94 |
| | V | 1.16 | 0.80 | 1.29 | 0.84 | 0.71 | 1.94 | 2.44 | 1.65 | 1.72 | 2.09 | 1.83 | 1.67 | 1.93 | 1.91 | 2.37 | 1.59 | 1.80 | 1.89 | 1.97 | 2.06 | 1.67 |
| | AVG | 0.02 | 0.44 | 0.66 | 0.58 | 0.62 | 0.05 | 1.10 | 1.89 | 1.39 | 0.91 | 1.14 | 1.53 | 1.10 | 1.00 | 1.27 | 0.66 | 1.18 | 0.90 | 0.85 | 1.15 | 0.82 |
| | MS | 2.28 | 3.27 | 1.91 | 2.69 | 0.85 | 0.19 | 1.14 | 2.38 | 1.85 | 1.40 | 1.72 | 2.04 | 1.68 | 1.55 | 2.56 | 1.62 | 2.04 | 2.08 | 2.09 | 2.29 | 1.83 |
| | MV | 0.86 | 0.38 | 0.20 | 1.40 | 1.65 | 0.82 | 1.89 | 1.73 | 2.06 | 2.74 | 2.34 | 1.94 | 2.48 | 2.49 | 1.72 | 2.12 | 1.70 | 1.81 | 1.91 | 1.68 | 1.88 |
| | SV | 0.31 | 2.84 | 0.80 | 1.41 | 1.84 | 0.47 | 2.00 | 1.58 | 1.35 | 1.75 | 1.48 | 1.38 | 1.64 | 1.50 | 2.10 | 1.14 | 1.52 | 1.51 | 1.57 | 1.76 | 1.27 |
| 2020 | M | 2.59 | 2.41 | 2.79 | 2.61 | 2.52 | 2.71 | 2.60 | 0.98 | 1.50 | 3.52 | 0.92 | 1.90 | 1.33 | 0.42 | 1.70 | 2.09 | 1.95 | 1.91 | 1.88 | 1.83 | 2.02 |
| | S | 1.79 | 1.77 | 1.73 | 1.75 | 1.77 | 1.76 | 1.74 | 1.75 | 2.50 | 2.31 | 0.53 | 0.74 | 0.56 | 1.28 | 0.96 | 1.14 | 1.23 | 0.99 | 0.95 | 1.03 | 1.09 |
| | V | 2.69 | 2.48 | 2.75 | 2.67 | 2.61 | 2.75 | 2.63 | 2.59 | 1.75 | 1.92 | 1.26 | 0.04 | 1.26 | 0.37 | 1.87 | 1.42 | 1.75 | 1.63 | 1.60 | 1.79 | 1.55 |
| | AVG | 2.18 | 1.80 | 2.35 | 2.08 | 1.99 | 2.26 | 2.01 | 2.59 | 1.04 | 0.92 | 1.02 | 1.00 | 0.18 | 1.01 | 1.33 | 1.16 | 1.68 | 1.32 | 1.16 | 1.48 | 1.38 |
| | MS | 2.51 | 2.12 | 2.68 | 2.43 | 2.33 | 2.61 | 2.36 | 2.27 | 0.50 | 1.27 | 1.16 | 0.29 | 0.31 | 0.53 | 0.65 | 1.45 | 1.48 | 1.10 | 0.95 | 1.03 | 1.40 |
| | MV | 2.44 | 2.08 | 2.61 | 2.37 | 2.27 | 2.53 | 2.31 | 2.18 | 0.74 | 0.30 | 1.16 | 1.69 | 2.08 | 0.33 | 1.14 | 1.39 | 1.58 | 1.29 | 1.18 | 1.33 | 1.43 |
| | SV | 2.24 | 1.91 | 2.37 | 2.16 | 2.08 | 2.30 | 2.09 | 1.79 | 0.81 | 0.75 | 1.28 | 0.99 | 0.59 | 0.54 | 1.51 | 1.87 | 1.66 | 1.64 | 1.66 | 1.56 | 1.73 |
| 2021 | M | 2.38 | 2.25 | 2.42 | 2.37 | 2.33 | 2.41 | 2.34 | 1.32 | 1.11 | 1.51 | 1.19 | 1.12 | 1.35 | 1.23 | 1.17 | 1.48 | 1.61 | 1.60 | 1.75 | 0.95 | 2.31 |
| | S | 1.27 | 0.85 | 1.42 | 1.08 | 1.02 | 1.29 | 1.00 | 1.77 | 1.02 | 1.69 | 1.35 | 1.23 | 1.67 | 1.29 | 2.12 | 3.07 | 1.97 | 1.44 | 0.55 | 2.48 | 0.85 |
| | V | 1.71 | 1.71 | 1.52 | 1.64 | 1.70 | 1.62 | 1.61 | 2.01 | 2.14 | 2.88 | 2.52 | 2.10 | 2.70 | 2.62 | 0.34 | 1.96 | 2.85 | 0.55 | 1.94 | 0.55 | 0.62 |
| | AVG | 2.07 | 1.97 | 2.23 | 2.08 | 2.02 | 2.14 | 2.08 | 2.36 | 0.98 | 1.98 | 2.01 | 2.03 | 2.12 | 1.90 | 1.73 | 1.82 | 0.57 | 1.28 | 1.01 | 1.53 | 1.42 |
| | MS | 2.69 | 2.27 | 2.66 | 2.56 | 2.50 | 2.72 | 2.45 | 2.33 | 1.75 | 1.60 | 1.77 | 1.96 | 1.82 | 1.62 | 0.49 | 2.24 | 1.29 | 0.29 | 1.64 | 0.56 | 1.52 |
| | MV | 1.31 | 1.21 | 1.24 | 1.20 | 1.23 | 1.25 | 1.16 | 1.34 | 0.97 | 2.25 | 1.51 | 1.03 | 1.95 | 1.61 | 1.46 | 1.12 | 1.39 | 0.65 | 0.42 | 1.24 | 1.65 |
| | SV | 1.26 | 1.15 | 1.36 | 1.18 | 1.17 | 1.26 | 1.17 | 1.68 | 0.96 | 2.17 | 1.55 | 1.16 | 2.01 | 1.56 | 1.22 | 0.75 | 2.17 | 0.64 | 1.96 | 1.52 | 0.89 |

| Trained on | Tested on | 2019 | | | | | | | 2020 | | | | | | | 2021 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | M | S | V | AVG | MS | MV | SV | M | S | V | AVG | MS | MV | SV | M | S | V | AVG | MS | MV | SV |
| 2019 | M | 2.06 | 2.06 | 2.89 | 2.50 | 2.06 | 1.64 | 1.26 | 1.96 | 1.14 | 1.98 | 1.60 | 1.39 | 1.97 | 1.53 | 1.49 | 1.74 | 1.69 | 1.59 | 1.58 | 1.57 | 1.68 |
| | S | 1.52 | 1.48 | 1.16 | 3.25 | 2.71 | 3.43 | 3.36 | 2.56 | 2.26 | 2.76 | 2.64 | 2.40 | 2.81 | 2.60 | 2.61 | 1.90 | 2.10 | 2.22 | 2.27 | 2.36 | 2.01 |
| | V | 1.91 | 1.64 | 1.34 | 1.19 | 1.23 | 1.11 | 2.87 | 2.23 | 1.56 | 2.60 | 2.17 | 1.78 | 2.56 | 2.14 | 1.46 | 1.32 | 1.13 | 1.18 | 1.32 | 1.20 | 1.12 |
| | AVG | 1.99 | 1.81 | 3.05 | 2.45 | 0.87 | 0.73 | 1.06 | 1.39 | 1.17 | 2.48 | 1.73 | 1.19 | 2.14 | 1.85 | 2.24 | 1.50 | 1.45 | 1.66 | 1.85 | 1.78 | 1.41 |
| | MS | 0.65 | 1.66 | 1.96 | 1.75 | 1.66 | 1.61 | 2.06 | 1.91 | 1.50 | 2.42 | 1.97 | 1.62 | 2.29 | 2.00 | 1.77 | 1.67 | 1.38 | 1.52 | 1.68 | 1.51 | 1.45 |
| | MV | 2.14 | 2.14 | 1.50 | 1.83 | 2.14 | 0.75 | 1.19 | 1.78 | 1.26 | 1.88 | 1.56 | 1.40 | 1.82 | 1.54 | 2.32 | 1.51 | 1.83 | 1.86 | 1.90 | 2.06 | 1.65 |
| | SV | 2.80 | 2.26 | 2.32 | 1.90 | 1.79 | 1.75 | 2.17 | 2.09 | 1.19 | 2.13 | 1.74 | 1.48 | 2.14 | 1.65 | 1.13 | 1.07 | 1.04 | 0.93 | 1.00 | 0.99 | 0.94 |
| 2020 | M | 1.95 | 1.55 | 1.99 | 1.79 | 1.74 | 1.96 | 1.70 | 2.73 | 2.04 | 2.11 | 1.29 | 1.96 | 2.07 | 2.11 | 2.09 | 1.31 | 1.71 | 1.65 | 1.65 | 1.87 | 1.47 |
| | S | 2.47 | 2.03 | 2.58 | 2.36 | 2.27 | 2.54 | 2.26 | 0.74 | 2.22 | 2.37 | 3.08 | 1.35 | 3.13 | 2.80 | 2.26 | 1.46 | 1.91 | 1.85 | 1.83 | 2.07 | 1.67 |
| | V | 2.11 | 2.26 | 2.17 | 2.19 | 2.19 | 2.14 | 2.24 | 2.28 | 2.38 | 2.33 | 1.08 | 0.79 | 1.21 | 1.49 | 2.58 | 2.29 | 2.37 | 2.47 | 2.48 | 2.51 | 2.37 |
| | AVG | 2.46 | 1.98 | 2.57 | 2.33 | 2.23 | 2.53 | 2.22 | 2.55 | 2.45 | 3.42 | 2.75 | 2.52 | 2.68 | 2.96 | 1.31 | 1.39 | 1.58 | 1.35 | 1.28 | 1.41 | 1.43 |
| | MS | 1.79 | 1.41 | 1.85 | 1.63 | 1.59 | 1.80 | 1.55 | 1.39 | 1.56 | 1.23 | 2.90 | 1.44 | 1.19 | 3.56 | 2.30 | 1.63 | 1.94 | 1.94 | 1.95 | 2.11 | 1.77 |
| | MV | 1.84 | 1.43 | 1.89 | 1.68 | 1.62 | 1.85 | 1.58 | 0.97 | 1.58 | 3.39 | 3.04 | 2.49 | 3.23 | 2.83 | 1.74 | 1.62 | 1.72 | 1.65 | 1.64 | 1.71 | 1.63 |
| | SV | 1.92 | 1.87 | 2.00 | 1.92 | 1.89 | 1.95 | 1.92 | 2.64 | 0.91 | 2.45 | 1.87 | 1.13 | 1.98 | 1.88 | 2.46 | 1.64 | 1.93 | 2.00 | 2.05 | 2.18 | 1.78 |
| 2021 | M | 0.83 | 0.77 | 0.76 | 0.70 | 0.75 | 0.74 | 0.67 | 1.25 | 1.31 | 2.34 | 1.68 | 1.22 | 1.98 | 1.84 | 1.41 | 2.45 | 2.58 | 2.39 | 2.28 | 2.86 | 1.89 |
| | S | 2.48 | 2.22 | 2.48 | 2.41 | 2.37 | 2.51 | 2.34 | 1.16 | 0.69 | 1.28 | 0.84 | 0.76 | 1.11 | 0.86 | 1.44 | 1.59 | 0.69 | 0.96 | 1.18 | 1.22 | 0.49 |
| | V | 2.56 | 2.13 | 2.67 | 2.46 | 2.36 | 2.64 | 2.36 | 1.86 | 1.29 | 1.79 | 1.55 | 1.45 | 1.78 | 1.49 | 2.58 | 1.53 | 1.76 | 0.84 | 0.87 | 1.24 | 0.77 |
| | AVG | 2.28 | 2.00 | 2.05 | 2.13 | 2.15 | 2.20 | 2.02 | 1.64 | 1.30 | 1.97 | 1.58 | 1.37 | 1.83 | 1.61 | 1.73 | 0.72 | 2.06 | 1.63 | 1.41 | 2.54 | 0.79 |
| | MS | 1.19 | 1.02 | 0.97 | 1.00 | 1.07 | 1.06 | 0.92 | 2.31 | 1.56 | 2.29 | 2.04 | 1.82 | 2.35 | 1.95 | 1.50 | 0.85 | 1.58 | 1.15 | 1.69 | 1.51 | 1.11 |
| | MV | 2.38 | 1.96 | 2.53 | 2.28 | 2.18 | 2.46 | 2.20 | 1.37 | 0.93 | 1.74 | 1.24 | 1.01 | 1.55 | 1.27 | 1.03 | 1.75 | 1.44 | 1.37 | 1.36 | 1.20 | 1.63 |
| | SV | 2.76 | 2.61 | 2.96 | 2.81 | 2.72 | 2.88 | 2.80 | 1.56 | 1.45 | 1.93 | 1.61 | 1.45 | 1.78 | 1.67 | 1.72 | 1.22 | 1.43 | 1.28 | 1.27 | 2.37 | 1.82 |

Figure A.8: Performance (SSE residual $r^2$) of the variety-specific Innovator model (top) and the variety-agnostic model tested on the Innovator variety (bottom). Rows correspond to the training dataset, columns show the testing configuration. SSE residuals $r^2 \geq 1$ are grayed out.



| | Tested on | 2019 | | | | | | | 2020 | | | | | | | 2021 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Trained on | | M | S | V | AVG | MS | MV | SV | M | S | V | AVG | MS | MV | SV | M | S | V | AVG | MS | MV | SV |
| 2019 | M | 0.79 | 0.83 | | 0.90 | | 0.75 | | | | | | | | | | 0.42 | | | | | |
| | S | | | | 0.74 | | | | | 0.49 | | | 0.41 | | | | 0.62 | | 0.45 | 0.47 | | 0.53 |
| | V | | | | | | | | | | | | | | | | | | | | | |
| | AVG | 0.99 | | 0.67 | | 0.69 | 0.97 | | | | 0.55 | 0.43 | | 0.45 | 0.50 | 0.37 | 0.67 | 0.41 | 0.55 | 0.58 | 0.43 | 0.59 |
| | MS | | | | | | 0.91 | | | | | | | | | | | | | | | |
| | MV | | 0.81 | 0.90 | | | | | | | -0.37 | | | | | | | | | | | |
| | SV | 0.85 | | | | | 0.76 | | | | | | | | | | 0.43 | | | | | 0.37 |
| 2020 | M | | | -0.40 | | | | | | | -0.76 | | | | 0.79 | | | | | | | |
| | S | | | | | | | | | | | 0.73 | | 0.72 | | 0.52 | 0.43 | 0.38 | 0.51 | 0.53 | 0.48 | 0.45 |
| | V | | | -0.38 | | | -0.38 | | | | | | 0.98 | | 0.81 | | | | | | | |
| | AVG | | | | | | | | | | | | | 0.91 | | | 0.42 | | | 0.42 | | |
| | MS | | | | | | | | | 0.75 | | | 0.86 | 0.84 | 0.73 | 0.67 | | | 0.45 | 0.52 | 0.49 | |
| | MV | | | | | | | | | | 0.85 | | | | 0.83 | 0.43 | | | | 0.41 | | |
| | SV | | | | | | | | | | | | | | 0.73 | | | | | | | |
| 2021 | M | | | | | | | | | 0.44 | | 0.41 | 0.44 | | 0.39 | | | | | | | |
| | S | 0.37 | 0.57 | | 0.46 | 0.49 | | 0.50 | | 0.49 | | | 0.39 | | | | | | | 0.73 | | |
| | V | | | | | | | | | | -0.44 | | | | | 0.83 | | | 0.72 | | 0.72 | 0.69 |
| | AVG | | | | | | | | | 0.51 | | | | | | | | 0.71 | | | | |
| | MS | | | | | | | | | | | | | | | | | | | | 0.72 | |
| | MV | | 0.40 | 0.38 | 0.40 | 0.38 | 0.38 | 0.42 | | 0.51 | | | 0.48 | | | | | | 0.68 | 0.79 | | |
| | SV | 0.37 | 0.43 | | 0.41 | 0.41 | 0.37 | 0.42 | | 0.52 | | | 0.42 | | | | | | 0.68 | | | |

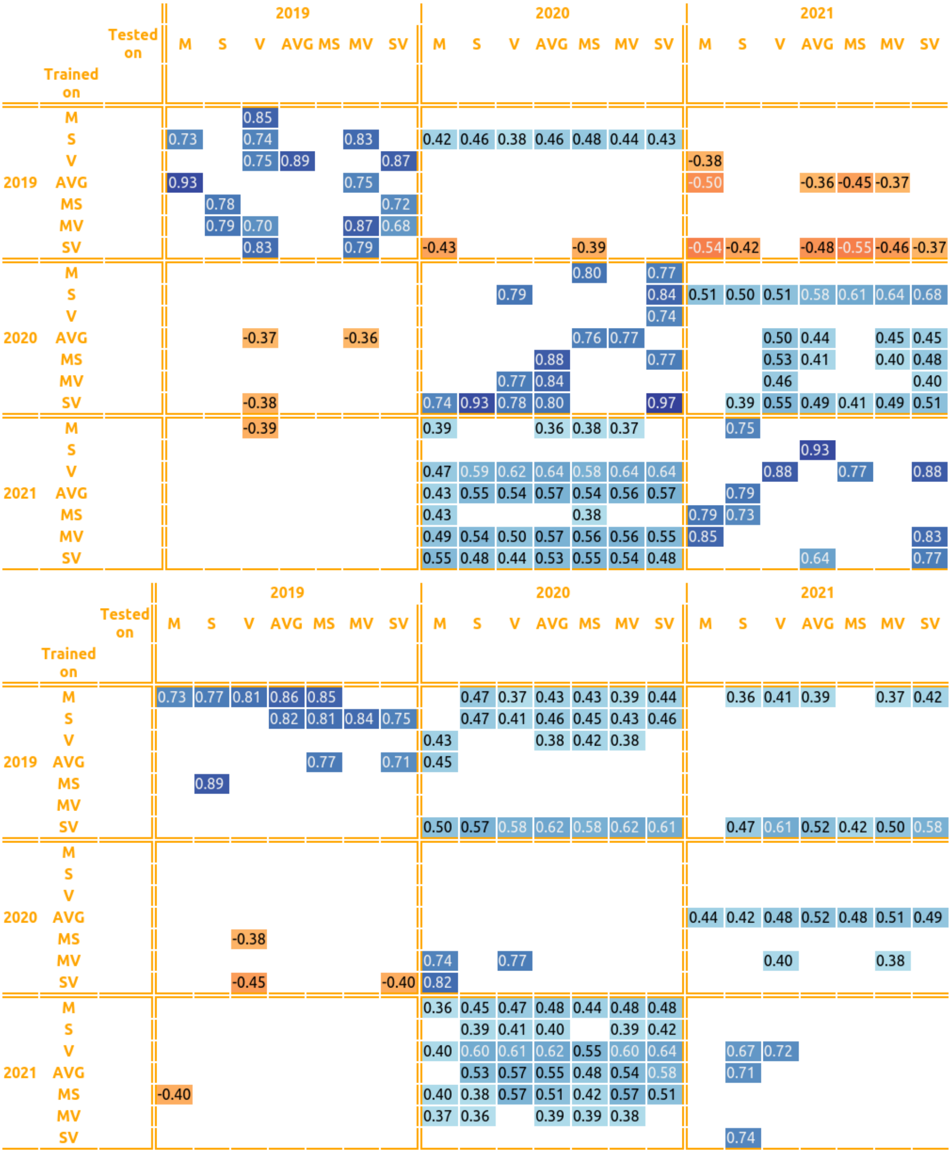
| | Tested on | 2019 | | | | | | | 2020 | | | | | | | 2021 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Trained on | | M | S | V | AVG | MS | MV | SV | M | S | V | AVG | MS | MV | SV | M | S | V | AVG | MS | MV | SV |
| 2019 | M | | | | | | | | | 0.43 | | | | | | | | | | | | |
| | S | | | | | | -0.72 | | | | -0.38 | | | -0.41 | | | | | | | | |
| | V | | | | | | | | | | | | | | | | | 0.44 | 0.41 | | 0.40 | 0.44 |
| | AVG | | | | | | | | | 0.42 | | | 0.41 | | | | | | | | | |
| | MS | | | | | | | | | | | | | | | | | | | | | |
| | MV | | | | | | | | | 0.37 | | | | | | | | | | | | |
| | SV | | | | | | | | | 0.40 | | | | | | 0.44 | 0.46 | 0.48 | 0.53 | 0.50 | 0.51 | 0.53 |
| 2020 | M | | | | | | | | | | | | | | | | | | | | | |
| | S | | | | | | | | | | | | | | | | | | | | | |
| | V | | | | | | | | | | | | | | | | | | | | | |
| | AVG | | | | | | | | | | -0.71 | | | | | | | | | 0.36 | | |
| | MS | | | | | | | | | | | | | | -0.78 | | | | | | | |
| | MV | | | | | | | | | | | | | | | | | | | | | |
| | SV | | | | | | | | | | | | | | | | | | | | | |
| 2021 | M | 0.59 | 0.61 | 0.62 | 0.65 | 0.62 | 0.63 | 0.66 | 0.38 | | | | 0.39 | | | | | | | | | |
| | S | | | | | | | | 0.42 | 0.65 | | 0.58 | 0.62 | 0.45 | 0.57 | | | | | | | 0.75 |
| | V | | | | | | | | | | | | | | | | | | | | | |
| | AVG | | | | | | | | | | | | | | | | | | | | | |
| | MS | 0.40 | 0.49 | 0.52 | 0.50 | 0.46 | 0.47 | 0.54 | | | | | | | | | | | | | | |
| | MV | | | | | | | | | 0.54 | | 0.38 | 0.50 | | 0.36 | | | | | | | |
| | SV | -0.38 | | -0.48 | -0.41 | | -0.44 | -0.40 | | | | | | | | | | | | | | |

Figure A.9: Performance (Pearson correlation $c$) of the variety-specific Innovator model (top) and the variety-agnostic model tested on the Innovator variety (bottom). Rows correspond to the training dataset, columns show the testing configuration. Correlations with $p \geq 0.05$ are omitted.

| | Tested on | 2019 | | | | | | | 2020 | | | | | | | 2021 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Trained on | | M | S | V | AVG | MS | MV | SV | M | S | V | AVG | MS | MV | SV | M | S | V | AVG | MS | MV | SV |
| 2019 | M | 1.02 | 1.89 | 0.29 | 0.95 | 1.01 | 1.04 | 1.00 | 1.42 | 1.30 | 1.47 | 1.34 | 1.30 | 1.39 | 1.37 | 2.03 | 1.93 | 1.79 | 1.86 | 1.91 | 1.90 | 1.77 |
| | S | 0.53 | 1.19 | 0.51 | 1.01 | 1.23 | 0.34 | 1.78 | 1.15 | 1.08 | 1.25 | 1.08 | 1.05 | 1.11 | 1.13 | 1.95 | 1.44 | 1.31 | 1.52 | 1.68 | 1.61 | 1.32 |
| | V | 1.72 | 1.29 | 0.50 | 0.23 | 1.84 | 1.64 | 0.26 | 1.92 | 1.75 | 1.46 | 1.63 | 1.81 | 1.58 | 1.57 | 2.76 | 2.44 | 2.07 | 2.50 | 2.69 | 2.49 | 2.27 |
| | AVG | 0.14 | 1.16 | 1.27 | 1.43 | 1.49 | 0.50 | 1.57 | 2.50 | 2.16 | 1.86 | 2.13 | 2.33 | 2.11 | 1.99 | 3.00 | 2.56 | 2.28 | 2.72 | 2.89 | 2.73 | 2.45 |
| | MS | 0.96 | 0.43 | 0.94 | 1.55 | 1.14 | 1.01 | 0.56 | 2.03 | 1.52 | 1.36 | 1.53 | 1.72 | 1.56 | 1.40 | 2.55 | 1.82 | 1.47 | 1.97 | 2.24 | 2.04 | 1.61 |
| | MV | 0.80 | 0.42 | 0.60 | 1.45 | 1.42 | 0.26 | 0.65 | 1.92 | 1.51 | 1.38 | 1.51 | 1.66 | 1.53 | 1.41 | 2.05 | 1.58 | 1.55 | 1.70 | 1.81 | 1.80 | 1.53 |
| | SV | 1.04 | 1.09 | 0.35 | 0.66 | 1.64 | 0.41 | 1.37 | 2.86 | 2.63 | 2.34 | 2.63 | 2.78 | 2.60 | 2.49 | 3.09 | 2.85 | 2.53 | 2.96 | 3.09 | 2.92 | 2.74 |
| 2020 | M | 1.77 | 2.16 | 2.62 | 2.28 | 1.97 | 2.29 | 2.48 | 1.92 | 1.32 | 0.96 | 0.77 | 0.40 | 1.54 | 0.45 | 1.74 | 1.28 | 1.48 | 1.44 | 1.48 | 1.57 | 1.33 |
| | S | 2.03 | 2.14 | 2.67 | 2.38 | 2.09 | 2.45 | 2.50 | 1.02 | 2.05 | 0.42 | 1.12 | 0.71 | 0.76 | 0.31 | 0.98 | 1.00 | 0.98 | 0.83 | 0.78 | 0.73 | 0.63 |
| | V | 1.72 | 1.83 | 1.87 | 1.78 | 1.75 | 1.79 | 1.83 | 1.89 | 1.92 | 1.27 | 0.87 | 1.11 | 1.10 | 0.51 | 2.00 | 1.48 | 1.63 | 1.67 | 1.73 | 1.80 | 1.52 |
| | AVG | 2.55 | 2.16 | 2.73 | 2.59 | 2.39 | 2.72 | 2.56 | 1.42 | 1.65 | 1.12 | 0.94 | 0.48 | 0.46 | 2.06 | 1.36 | 1.38 | 0.99 | 1.13 | 1.29 | 1.09 | 1.11 |
| | MS | 2.25 | 2.07 | 2.51 | 2.36 | 2.18 | 2.45 | 2.37 | 0.84 | 1.43 | 1.62 | 0.24 | 1.95 | 0.67 | 0.46 | 1.60 | 1.29 | 0.94 | 1.18 | 1.39 | 1.21 | 1.03 |
| | MV | 1.85 | 2.07 | 1.88 | 1.91 | 1.96 | 1.85 | 1.95 | 0.91 | 1.45 | 0.46 | 0.33 | 0.74 | 0.76 | 2.49 | 1.76 | 1.45 | 1.07 | 1.36 | 1.57 | 1.37 | 1.20 |
| | SV | 2.31 | 2.17 | 2.75 | 2.53 | 2.27 | 2.63 | 2.58 | 0.53 | 0.14 | 0.43 | 0.40 | 0.69 | 1.51 | 0.07 | 1.32 | 1.22 | 0.90 | 1.02 | 1.18 | 1.02 | 0.97 |
| 2021 | M | 2.27 | 2.00 | 2.78 | 2.47 | 2.15 | 2.63 | 2.51 | 1.21 | 1.34 | 1.43 | 1.27 | 1.23 | 1.27 | 1.36 | 1.92 | 0.50 | 0.76 | 2.86 | 1.01 | 1.61 | 2.32 |
| | S | 1.88 | 1.89 | 2.33 | 2.03 | 1.87 | 2.16 | 2.16 | 1.96 | 1.93 | 1.77 | 1.86 | 1.94 | 1.82 | 1.84 | 1.03 | 0.34 | 0.76 | 0.14 | 1.18 | 0.90 | 1.39 |
| | V | 2.02 | 1.96 | 1.97 | 1.98 | 1.99 | 1.99 | 1.96 | 1.06 | 0.82 | 0.76 | 0.72 | 0.84 | 0.73 | 0.72 | 2.80 | 2.78 | 0.24 | 1.92 | 0.46 | 0.88 | 0.25 |
| | AVG | 1.76 | 1.74 | 1.98 | 1.82 | 1.72 | 1.88 | 1.86 | 1.15 | 0.90 | 0.92 | 0.85 | 0.93 | 0.88 | 0.85 | 1.66 | 0.42 | 1.71 | 1.92 | 0.96 | 2.31 | 1.39 |
| | MS | 1.96 | 2.00 | 2.52 | 2.24 | 1.98 | 2.32 | 2.34 | 1.13 | 1.41 | 1.67 | 1.39 | 1.24 | 1.41 | 1.54 | 0.42 | 0.54 | 1.07 | 1.18 | 1.57 | 0.86 | 1.73 |
| | MV | 1.81 | 1.89 | 2.19 | 1.99 | 1.84 | 2.04 | 2.07 | 1.01 | 0.92 | 0.99 | 0.86 | 0.88 | 0.87 | 0.91 | 0.30 | 1.34 | 0.99 | 1.74 | 1.47 | 2.50 | 0.34 |
| | SV | 2.03 | 1.94 | 2.53 | 2.24 | 1.98 | 2.36 | 2.32 | 0.91 | 1.03 | 1.12 | 0.93 | 0.90 | 0.92 | 1.04 | 1.11 | 2.50 | 0.79 | 0.73 | 1.22 | 0.81 | 0.46 |

| | Tested on | 2019 | | | | | | | 2020 | | | | | | | 2021 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Trained on | | M | S | V | AVG | MS | MV | SV | M | S | V | AVG | MS | MV | SV | M | S | V | AVG | MS | MV | SV |
| 2019 | M | 0.53 | 0.46 | 0.38 | 0.29 | 0.29 | 0.82 | 0.89 | 1.39 | 1.06 | 1.25 | 1.13 | 1.14 | 1.22 | 1.13 | 1.49 | 1.28 | 1.17 | 1.22 | 1.32 | 1.26 | 1.16 |
| | S | 2.26 | 2.71 | 2.61 | 0.37 | 0.37 | 0.33 | 0.49 | 1.30 | 1.06 | 1.19 | 1.08 | 1.10 | 1.13 | 1.08 | 2.44 | 2.01 | 1.75 | 2.09 | 2.27 | 2.12 | 1.87 |
| | V | 1.64 | 2.31 | 2.63 | 2.62 | 2.61 | 2.74 | 1.37 | 1.14 | 1.29 | 1.43 | 1.23 | 1.17 | 1.24 | 1.33 | 1.53 | 1.33 | 1.63 | 1.43 | 1.38 | 1.53 | 1.44 |
| | AVG | 1.60 | 2.99 | 1.69 | 1.88 | 0.45 | 0.79 | 0.58 | 1.09 | 1.66 | 1.82 | 1.54 | 1.38 | 1.50 | 1.74 | 2.11 | 1.76 | 2.24 | 2.05 | 1.94 | 2.19 | 2.01 |
| | MS | 0.88 | 0.22 | 0.89 | 0.91 | 0.97 | 0.70 | 1.22 | 1.40 | 1.85 | 2.09 | 1.82 | 1.63 | 1.81 | 1.98 | 1.53 | 1.66 | 1.93 | 1.66 | 1.54 | 1.69 | 1.78 |
| | MV | 2.76 | 2.62 | 2.87 | 2.77 | 2.72 | 1.04 | 1.05 | 1.54 | 1.33 | 1.34 | 1.32 | 1.38 | 1.34 | 1.30 | 2.20 | 1.71 | 1.97 | 1.97 | 1.97 | 2.11 | 1.83 |
| | SV | 0.67 | 0.92 | 1.30 | 0.67 | 0.65 | 0.79 | 0.61 | 0.99 | 0.85 | 0.84 | 0.75 | 0.83 | 0.76 | 0.79 | 1.40 | 1.07 | 0.79 | 0.96 | 1.16 | 1.00 | 0.83 |
| 2020 | M | 1.66 | 2.21 | 2.29 | 2.10 | 1.94 | 2.03 | 2.29 | 3.38 | 3.03 | 3.10 | 1.28 | 2.64 | 2.77 | 3.15 | 1.90 | 1.71 | 1.84 | 1.80 | 1.79 | 1.86 | 1.76 |
| | S | 1.80 | 2.32 | 2.38 | 2.23 | 2.08 | 2.15 | 2.41 | 1.27 | 1.15 | 1.08 | 0.80 | 1.18 | 0.85 | 0.71 | 2.12 | 1.77 | 1.81 | 1.90 | 1.95 | 1.97 | 1.77 |
| | V | 1.86 | 2.21 | 2.31 | 2.18 | 2.05 | 2.14 | 2.30 | 1.89 | 2.74 | 1.38 | 1.51 | 1.62 | 1.46 | 1.47 | 2.28 | 2.18 | 1.94 | 2.16 | 2.26 | 2.14 | 2.06 |
| | AVG | 1.93 | 2.36 | 2.69 | 2.44 | 2.17 | 2.41 | 2.63 | 1.95 | 2.53 | 2.26 | 2.02 | 2.19 | 2.01 | 2.33 | 1.12 | 1.15 | 1.04 | 0.97 | 1.04 | 0.98 | 1.02 |
| | MS | 2.21 | 2.12 | 2.76 | 2.48 | 2.18 | 2.59 | 2.55 | 1.51 | 1.55 | 1.11 | 0.92 | 1.52 | 1.28 | 1.05 | 1.83 | 1.65 | 1.28 | 1.53 | 1.72 | 1.52 | 1.41 |
| | MV | 2.42 | 2.46 | 2.70 | 2.65 | 2.50 | 2.64 | 2.68 | 0.51 | 0.68 | 0.45 | 1.07 | 1.28 | 0.82 | 1.35 | 1.43 | 1.50 | 1.19 | 1.28 | 1.40 | 1.24 | 1.28 |
| | SV | 2.06 | 2.43 | 2.89 | 2.61 | 2.29 | 2.61 | 2.80 | 0.35 | 1.56 | 0.97 | 0.76 | 1.81 | 1.33 | 1.11 | 1.93 | 2.14 | 2.43 | 2.18 | 2.03 | 2.19 | 2.32 |
| 2021 | M | 2.44 | 2.14 | 2.54 | 2.46 | 2.32 | 2.55 | 2.42 | 1.27 | 1.11 | 1.07 | 1.03 | 1.12 | 1.04 | 1.04 | 2.94 | 0.96 | 2.62 | 2.53 | 2.41 | 1.75 | 2.73 |
| | S | 2.03 | 2.23 | 1.72 | 1.96 | 2.15 | 1.83 | 1.93 | 1.52 | 1.21 | 1.19 | 1.20 | 1.29 | 1.22 | 1.16 | 2.28 | 1.89 | 3.02 | 2.66 | 2.12 | 1.65 | 2.41 |
| | V | 1.45 | 1.90 | 1.59 | 1.58 | 1.64 | 1.49 | 1.68 | 1.21 | 0.79 | 0.78 | 0.77 | 0.89 | 0.81 | 0.73 | 2.37 | 0.65 | 0.56 | 1.07 | 1.16 | 1.39 | 1.04 |
| | AVG | 1.96 | 1.84 | 1.47 | 1.67 | 1.88 | 1.64 | 1.57 | 1.34 | 0.95 | 0.86 | 0.90 | 1.05 | 0.92 | 0.84 | 1.82 | 0.59 | 0.69 | 0.91 | 1.12 | 1.93 | 0.98 |
| | MS | 2.81 | 2.00 | 2.22 | 2.38 | 2.44 | 2.51 | 2.14 | 1.21 | 1.24 | 0.87 | 0.97 | 1.17 | 0.87 | 0.98 | 2.32 | 1.19 | 1.62 | 1.22 | 2.29 | 2.06 | 2.08 |
| | MV | 2.17 | 2.21 | 2.21 | 2.24 | 2.21 | 2.21 | 2.24 | 1.26 | 1.28 | 1.35 | 1.23 | 1.22 | 1.23 | 1.28 | 1.29 | 1.91 | 1.66 | 1.57 | 1.58 | 1.35 | 1.90 |
| | SV | 1.63 | 2.26 | 2.23 | 2.08 | 1.95 | 1.98 | 2.28 | 1.88 | 2.01 | 2.40 | 2.16 | 1.95 | 2.24 | 2.24 | 2.70 | 0.53 | 0.86 | 0.71 | 0.76 | 1.02 | 2.93 |

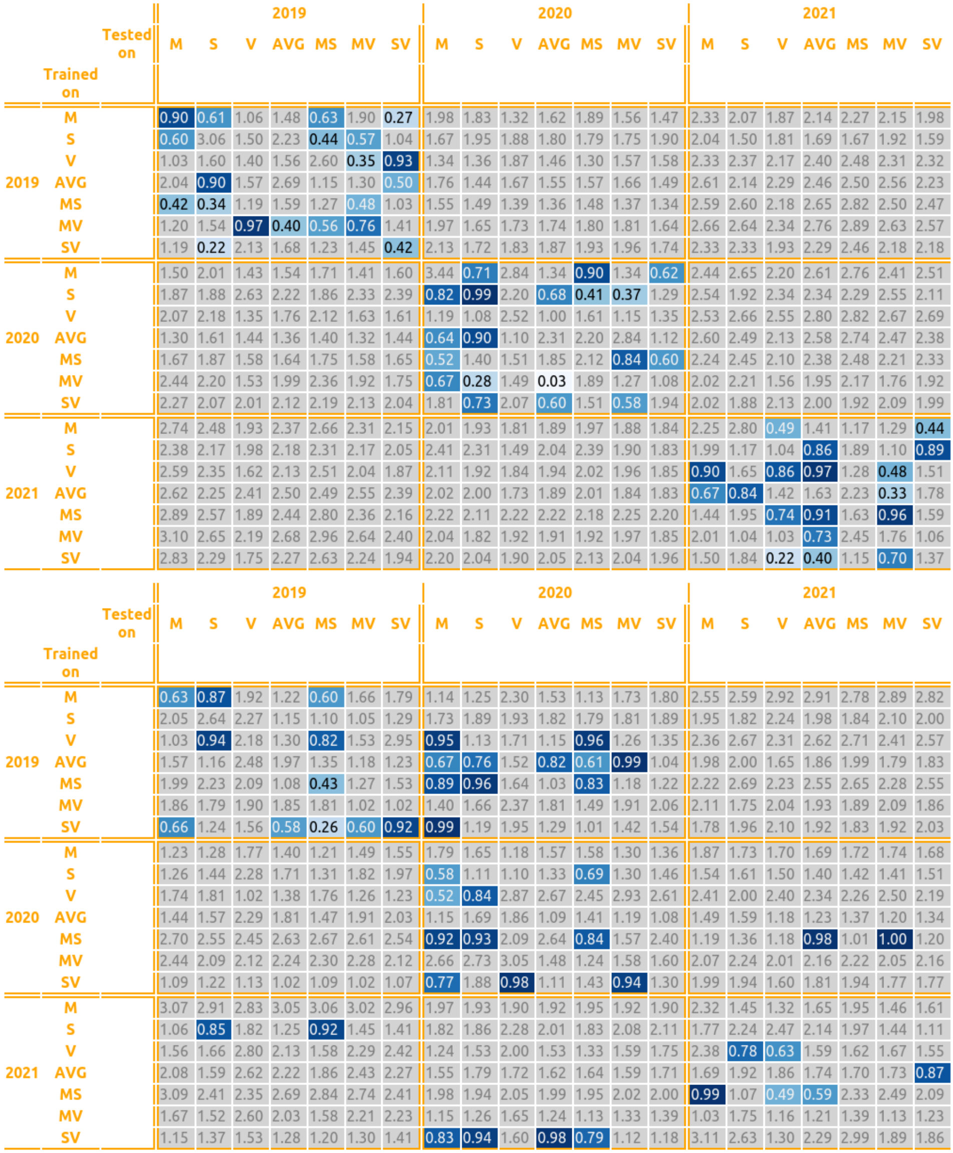
Figure A.10: Performance (SSE residual $r^2$) of the variety-specific Sagitta model (top) and the variety-agnostic model tested on the Sagitta variety (bottom). Rows correspond to the training dataset, columns show the testing configuration. SSE residuals $r^2 \geq 1$ are grayed out.



| | Tested on | 2019 | | | | | | | 2020 | | | | | | | 2021 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Trained on | | M | S | V | AVG | MS | MV | SV | M | S | V | AVG | MS | MV | SV | M | S | V | AVG | MS | MV | SV |
| 2019 | M | | | 0.85 | | | | | | | | | | | | | | | | | | |
| | S | 0.73 | | 0.74 | | | 0.83 | | 0.42 | 0.46 | 0.38 | 0.46 | 0.48 | 0.44 | 0.43 | | | | | | | |
| | V | | | 0.75 | 0.89 | | | 0.87 | | | | | | | | -0.38 | | | | | | |
| | AVG | 0.93 | | | | | 0.75 | | | | | | | | | -0.50 | | | -0.36 | -0.45 | -0.37 | |
| | MS | | 0.78 | | | | | 0.72 | | | | | | | | | | | | | | |
| | MV | | 0.79 | 0.70 | | | 0.87 | 0.68 | | | | | | | | | | | | | | |
| | SV | | | 0.83 | | | 0.79 | | -0.43 | | | | -0.39 | | | -0.54 | -0.42 | | -0.48 | -0.55 | -0.46 | -0.37 |
| 2020 | M | | | | | | | | | | | | 0.80 | | 0.77 | | | | | | | |
| | S | | | | | | | | | | 0.79 | | | | 0.84 | 0.51 | 0.50 | 0.51 | 0.58 | 0.61 | 0.64 | 0.68 |
| | V | | | | | | | | | | | | | | 0.74 | | | | | | | |
| | AVG | | | -0.37 | | | -0.36 | | | | | | 0.76 | 0.77 | | | | 0.50 | 0.44 | | 0.45 | 0.45 |
| | MS | | | | | | | | | | | 0.88 | | | 0.77 | | | 0.53 | 0.41 | | 0.40 | 0.48 |
| | MV | | | | | | | | | | 0.77 | 0.84 | | | | | | 0.46 | | | | 0.40 |
| | SV | | | -0.38 | | | | | 0.74 | 0.93 | 0.78 | 0.80 | | | 0.97 | | 0.39 | 0.55 | 0.49 | 0.41 | 0.49 | 0.51 |
| 2021 | M | | | -0.39 | | | | | 0.39 | | | 0.36 | 0.38 | 0.37 | | | 0.75 | | | | | |
| | S | | | | | | | | | | | | | | | | | | 0.93 | | | |
| | V | | | | | | | | 0.47 | 0.59 | 0.62 | 0.64 | 0.58 | 0.64 | 0.64 | | | 0.88 | | 0.77 | | 0.88 |
| | AVG | | | | | | | | 0.43 | 0.55 | 0.54 | 0.57 | 0.54 | 0.56 | 0.57 | | 0.79 | | | | | |
| | MS | | | | | | | | 0.43 | | | | 0.38 | | | 0.79 | 0.73 | | | | | |
| | MV | | | | | | | | 0.49 | 0.54 | 0.50 | 0.57 | 0.56 | 0.56 | 0.55 | 0.85 | | | | | | 0.83 |
| | SV | | | | | | | | 0.55 | 0.48 | 0.44 | 0.53 | 0.55 | 0.54 | 0.48 | | | | 0.64 | | | 0.77 |

| | Tested on | 2019 | | | | | | | 2020 | | | | | | | 2021 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Trained on | | M | S | V | AVG | MS | MV | SV | M | S | V | AVG | MS | MV | SV | M | S | V | AVG | MS | MV | SV |
| 2019 | M | 0.73 | 0.77 | 0.81 | 0.86 | 0.85 | | | | 0.47 | 0.37 | 0.43 | 0.43 | 0.39 | 0.44 | | 0.36 | 0.41 | 0.39 | | 0.37 | 0.42 |
| | S | | | | 0.82 | 0.81 | 0.84 | 0.75 | | 0.47 | 0.41 | 0.46 | 0.45 | 0.43 | 0.46 | | | | | | | |
| | V | | | | | | | | 0.43 | | | 0.38 | 0.42 | 0.38 | | | | | | | | |
| | AVG | | | | | 0.77 | | 0.71 | 0.45 | | | | | | | | | | | | | |
| | MS | | 0.89 | | | | | | | | | | | | | | | | | | | |
| | MV | | | | | | | | | | | | | | | | | | | | | |
| | SV | | | | | | | | 0.50 | 0.57 | 0.58 | 0.62 | 0.58 | 0.62 | 0.61 | | 0.47 | 0.61 | 0.52 | 0.42 | 0.50 | 0.58 |
| 2020 | M | | | | | | | | | | | | | | | | | | | | | |
| | S | | | | | | | | | | | | | | | | | | | | | |
| | V | | | | | | | | | | | | | | | | | | | | | |
| | AVG | | | | | | | | | | | | | | | 0.44 | 0.42 | 0.48 | 0.52 | 0.48 | 0.51 | 0.49 |
| | MS | | | -0.38 | | | | | | | | | | | | | | | | | | |
| | MV | | | | | | | | 0.74 | | 0.77 | | | | | | | 0.40 | | | 0.38 | |
| | SV | | | -0.45 | | | | -0.40 | 0.82 | | | | | | | | | | | | | |
| 2021 | M | | | | | | | | 0.36 | 0.45 | 0.47 | 0.48 | 0.44 | 0.48 | 0.48 | | | | | | | |
| | S | | | | | | | | | 0.39 | 0.41 | 0.40 | | 0.39 | 0.42 | | | | | | | |
| | V | | | | | | | | 0.40 | 0.60 | 0.61 | 0.62 | 0.55 | 0.60 | 0.64 | | 0.67 | 0.72 | | | | |
| | AVG | | | | | | | | | 0.53 | 0.57 | 0.55 | 0.48 | 0.54 | 0.58 | | 0.71 | | | | | |
| | MS | -0.40 | | | | | | | 0.40 | 0.38 | 0.57 | 0.51 | 0.42 | 0.57 | 0.51 | | | | | | | |
| | MV | | | | | | | | 0.37 | 0.36 | | 0.39 | 0.39 | 0.38 | | | | | | | | |
| | SV | | | | | | | | | | | | | | | | 0.74 | | | | | |

Figure A.11: Performance (Pearson correlation $c$) of the variety-specific Sagitta model (top) and the variety-agnostic model tested on the Sagitta variety (bottom). Rows correspond to the training dataset, columns show the testing configuration. Correlations with $p \geq 0.05$ are omitted.



| Trained on | Tested on | 2019 | | | | | | | 2020 | | | | | | | 2021 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | M | S | V | AVG | MS | MV | SV | M | S | V | AVG | MS | MV | SV | M | S | V | AVG | MS | MV | SV |
| | M | 0.90 | 0.61 | 1.06 | 1.48 | 0.63 | 1.90 | 0.27 | 1.98 | 1.83 | 1.32 | 1.62 | 1.89 | 1.56 | 1.47 | 2.33 | 2.07 | 1.87 | 2.14 | 2.27 | 2.15 | 1.98 |
| | S | 0.60 | 3.06 | 1.50 | 2.23 | 0.44 | 0.57 | 1.04 | 1.67 | 1.95 | 1.88 | 1.80 | 1.79 | 1.75 | 1.90 | 2.04 | 1.50 | 1.81 | 1.69 | 1.67 | 1.92 | 1.59 |
| | V | 1.03 | 1.60 | 1.40 | 1.56 | 2.60 | 0.35 | 0.93 | 1.34 | 1.36 | 1.87 | 1.46 | 1.30 | 1.57 | 1.58 | 2.33 | 2.37 | 2.17 | 2.40 | 2.48 | 2.31 | 2.32 |
| 2019 | AVG | 2.04 | 0.90 | 1.57 | 2.69 | 1.15 | 1.30 | 0.50 | 1.76 | 1.44 | 1.67 | 1.55 | 1.57 | 1.66 | 1.49 | 2.61 | 2.14 | 2.29 | 2.46 | 2.50 | 2.56 | 2.23 |
| | MS | 0.42 | 0.34 | 1.19 | 1.59 | 1.27 | 0.48 | 1.03 | 1.55 | 1.49 | 1.39 | 1.36 | 1.48 | 1.37 | 1.34 | 2.59 | 2.60 | 2.18 | 2.65 | 2.82 | 2.50 | 2.47 |
| | MV | 1.20 | 1.54 | 0.97 | 0.40 | 0.56 | 0.76 | 1.41 | 1.97 | 1.65 | 1.73 | 1.74 | 1.80 | 1.81 | 1.64 | 2.66 | 2.64 | 2.34 | 2.76 | 2.89 | 2.63 | 2.57 |
| | SV | 1.19 | 0.22 | 2.13 | 1.68 | 1.23 | 1.45 | 0.42 | 2.13 | 1.72 | 1.83 | 1.87 | 1.93 | 1.96 | 1.74 | 2.33 | 2.33 | 1.93 | 2.29 | 2.46 | 2.18 | 2.18 |
| | M | 1.50 | 2.01 | 1.43 | 1.54 | 1.71 | 1.41 | 1.60 | 3.44 | 0.71 | 2.84 | 1.34 | 0.90 | 1.34 | 0.62 | 2.44 | 2.65 | 2.20 | 2.61 | 2.76 | 2.41 | 2.51 |
| | S | 1.87 | 1.88 | 2.63 | 2.22 | 1.86 | 2.33 | 2.39 | 0.82 | 0.99 | 2.20 | 0.68 | 0.41 | 0.37 | 1.29 | 2.54 | 1.92 | 2.34 | 2.34 | 2.29 | 2.55 | 2.11 |
| | V | 2.07 | 2.18 | 1.35 | 1.76 | 2.12 | 1.63 | 1.61 | 1.19 | 1.08 | 2.52 | 1.00 | 1.61 | 1.15 | 1.35 | 2.53 | 2.66 | 2.55 | 2.80 | 2.82 | 2.67 | 2.69 |
| 2020 | AVG | 1.30 | 1.61 | 1.44 | 1.36 | 1.40 | 1.32 | 1.44 | 0.64 | 0.90 | 1.10 | 2.31 | 2.20 | 2.84 | 1.12 | 2.60 | 2.49 | 2.13 | 2.58 | 2.74 | 2.47 | 2.38 |
| | MS | 1.67 | 1.87 | 1.58 | 1.64 | 1.75 | 1.58 | 1.65 | 0.52 | 1.40 | 1.51 | 1.85 | 2.12 | 0.84 | 0.60 | 2.24 | 2.45 | 2.10 | 2.38 | 2.48 | 2.21 | 2.33 |
| | MV | 2.44 | 2.20 | 1.53 | 1.99 | 2.36 | 1.92 | 1.75 | 0.67 | 0.28 | 1.49 | 0.03 | 1.89 | 1.27 | 1.08 | 2.02 | 2.21 | 1.56 | 1.95 | 2.17 | 1.76 | 1.92 |
| | SV | 2.27 | 2.07 | 2.01 | 2.12 | 2.19 | 2.13 | 2.04 | 1.81 | 0.73 | 2.07 | 0.60 | 1.51 | 0.58 | 1.94 | 2.02 | 1.88 | 2.13 | 2.00 | 1.92 | 2.09 | 1.99 |
| | M | 2.74 | 2.48 | 1.93 | 2.37 | 2.66 | 2.31 | 2.15 | 2.01 | 1.93 | 1.81 | 1.89 | 1.97 | 1.88 | 1.84 | 2.25 | 2.80 | 0.49 | 1.41 | 1.17 | 1.29 | 0.44 |
| | S | 2.38 | 2.17 | 1.98 | 2.18 | 2.31 | 2.17 | 2.05 | 2.41 | 2.31 | 1.49 | 2.04 | 2.39 | 1.90 | 1.83 | 1.99 | 1.17 | 1.04 | 0.86 | 1.89 | 1.10 | 0.89 |
| | V | 2.59 | 2.35 | 1.62 | 2.13 | 2.51 | 2.04 | 1.87 | 2.11 | 1.92 | 1.84 | 1.94 | 2.02 | 1.96 | 1.85 | 0.90 | 1.65 | 0.86 | 0.97 | 1.28 | 0.48 | 1.51 |
| 2021 | AVG | 2.62 | 2.25 | 2.41 | 2.50 | 2.49 | 2.55 | 2.39 | 2.02 | 2.00 | 1.73 | 1.89 | 2.01 | 1.84 | 1.83 | 0.67 | 0.84 | 1.42 | 1.63 | 2.23 | 0.33 | 1.78 |
| | MS | 2.89 | 2.57 | 1.89 | 2.44 | 2.80 | 2.36 | 2.16 | 2.22 | 2.11 | 2.22 | 2.22 | 2.18 | 2.25 | 2.20 | 1.44 | 1.95 | 0.74 | 0.91 | 1.63 | 0.96 | 1.59 |
| | MV | 3.10 | 2.65 | 2.19 | 2.68 | 2.96 | 2.64 | 2.40 | 2.04 | 1.82 | 1.92 | 1.91 | 1.92 | 1.97 | 1.85 | 2.01 | 1.04 | 1.03 | 0.73 | 2.45 | 1.76 | 1.06 |
| | SV | 2.83 | 2.29 | 1.75 | 2.27 | 2.63 | 2.24 | 1.94 | 2.20 | 2.04 | 1.90 | 2.05 | 2.13 | 2.04 | 1.96 | 1.50 | 1.84 | 0.22 | 0.40 | 1.15 | 0.70 | 1.37 |

| Trained on | Tested on | 2019 | | | | | | | 2020 | | | | | | | 2021 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | M | S | V | AVG | MS | MV | SV | M | S | V | AVG | MS | MV | SV | M | S | V | AVG | MS | MV | SV |
| | M | 0.63 | 0.87 | 1.92 | 1.22 | 0.60 | 1.66 | 1.79 | 1.14 | 1.25 | 2.30 | 1.53 | 1.13 | 1.73 | 1.80 | 2.55 | 2.59 | 2.92 | 2.91 | 2.78 | 2.89 | 2.82 |
| | S | 2.05 | 2.64 | 2.27 | 1.15 | 1.10 | 1.05 | 1.29 | 1.73 | 1.89 | 1.93 | 1.82 | 1.79 | 1.81 | 1.89 | 1.95 | 1.82 | 2.24 | 1.98 | 1.84 | 2.10 | 2.00 |
| | V | 1.03 | 0.94 | 2.18 | 1.30 | 0.82 | 1.53 | 2.95 | 0.95 | 1.13 | 1.71 | 1.15 | 0.96 | 1.26 | 1.35 | 2.36 | 2.67 | 2.31 | 2.62 | 2.71 | 2.41 | 2.57 |
| 2019 | AVG | 1.57 | 1.16 | 2.48 | 1.97 | 1.35 | 1.18 | 1.23 | 0.67 | 0.76 | 1.52 | 0.82 | 0.61 | 0.99 | 1.04 | 1.98 | 2.00 | 1.65 | 1.86 | 1.99 | 1.79 | 1.83 |
| | MS | 1.99 | 2.23 | 2.09 | 1.08 | 0.43 | 1.27 | 1.53 | 0.89 | 0.96 | 1.64 | 1.03 | 0.83 | 1.18 | 1.22 | 2.22 | 2.69 | 2.23 | 2.55 | 2.65 | 2.28 | 2.55 |
| | MV | 1.86 | 1.79 | 1.90 | 1.85 | 1.81 | 1.02 | 1.02 | 1.40 | 1.66 | 2.37 | 1.81 | 1.49 | 1.91 | 2.06 | 2.11 | 1.75 | 2.04 | 1.93 | 1.89 | 2.09 | 1.86 |
| | SV | 0.66 | 1.24 | 1.56 | 0.58 | 0.26 | 0.60 | 0.92 | 0.99 | 1.19 | 1.95 | 1.29 | 1.01 | 1.42 | 1.54 | 1.78 | 1.96 | 2.10 | 1.92 | 1.83 | 1.92 | 2.03 |
| | M | 1.23 | 1.28 | 1.77 | 1.40 | 1.21 | 1.49 | 1.55 | 1.79 | 1.65 | 1.18 | 1.57 | 1.58 | 1.30 | 1.36 | 1.87 | 1.73 | 1.70 | 1.69 | 1.72 | 1.74 | 1.68 |
| | S | 1.26 | 1.44 | 2.28 | 1.71 | 1.31 | 1.82 | 1.97 | 0.58 | 1.11 | 1.10 | 1.33 | 0.69 | 1.30 | 1.46 | 1.54 | 1.61 | 1.50 | 1.40 | 1.42 | 1.41 | 1.51 |
| | V | 1.74 | 1.81 | 1.02 | 1.38 | 1.76 | 1.26 | 1.23 | 0.52 | 0.84 | 2.87 | 2.67 | 2.45 | 2.93 | 2.61 | 2.41 | 2.00 | 2.40 | 2.34 | 2.26 | 2.50 | 2.19 |
| 2020 | AVG | 1.44 | 1.57 | 2.29 | 1.81 | 1.47 | 1.91 | 2.03 | 1.15 | 1.69 | 1.86 | 1.09 | 1.41 | 1.19 | 1.08 | 1.49 | 1.59 | 1.18 | 1.23 | 1.37 | 1.20 | 1.34 |
| | MS | 2.70 | 2.55 | 2.45 | 2.63 | 2.67 | 2.61 | 2.54 | 0.92 | 0.93 | 2.09 | 2.64 | 0.84 | 1.57 | 2.40 | 1.19 | 1.36 | 1.18 | 0.98 | 1.01 | 1.00 | 1.20 |
| | MV | 2.44 | 2.09 | 2.12 | 2.24 | 2.30 | 2.28 | 2.12 | 2.66 | 2.73 | 3.05 | 1.48 | 1.24 | 1.58 | 1.60 | 2.07 | 2.24 | 2.01 | 2.16 | 2.22 | 2.05 | 2.16 |
| | SV | 1.09 | 1.22 | 1.13 | 1.02 | 1.09 | 1.02 | 1.07 | 0.77 | 1.88 | 0.98 | 1.11 | 1.43 | 0.94 | 1.30 | 1.99 | 1.94 | 1.60 | 1.81 | 1.94 | 1.77 | 1.77 |
| | M | 3.07 | 2.91 | 2.83 | 3.05 | 3.06 | 3.02 | 2.96 | 1.97 | 1.93 | 1.90 | 1.92 | 1.95 | 1.92 | 1.90 | 2.32 | 1.45 | 1.32 | 1.65 | 1.95 | 1.46 | 1.61 |
| | S | 1.06 | 0.85 | 1.82 | 1.25 | 0.92 | 1.45 | 1.41 | 1.82 | 1.86 | 2.28 | 2.01 | 1.83 | 2.08 | 2.11 | 1.77 | 2.24 | 2.47 | 2.14 | 1.97 | 1.44 | 1.11 |
| | V | 1.56 | 1.66 | 2.80 | 2.13 | 1.58 | 2.29 | 2.42 | 1.24 | 1.53 | 2.00 | 1.53 | 1.33 | 1.59 | 1.75 | 2.38 | 0.78 | 0.63 | 1.59 | 1.62 | 1.67 | 1.55 |
| 2021 | AVG | 2.08 | 1.59 | 2.62 | 2.22 | 1.86 | 2.43 | 2.27 | 1.55 | 1.79 | 1.72 | 1.62 | 1.64 | 1.59 | 1.71 | 1.69 | 1.92 | 1.86 | 1.74 | 1.70 | 1.73 | 0.87 |
| | MS | 3.09 | 2.41 | 2.35 | 2.69 | 2.84 | 2.74 | 2.41 | 1.98 | 1.94 | 2.05 | 1.99 | 1.95 | 2.02 | 2.00 | 0.99 | 1.07 | 0.49 | 0.59 | 2.33 | 2.49 | 2.09 |
| | MV | 1.67 | 1.52 | 2.60 | 2.03 | 1.58 | 2.21 | 2.23 | 1.15 | 1.26 | 1.65 | 1.24 | 1.13 | 1.33 | 1.39 | 1.03 | 1.75 | 1.16 | 1.21 | 1.39 | 1.13 | 1.23 |
| | SV | 1.15 | 1.37 | 1.53 | 1.28 | 1.20 | 1.30 | 1.41 | 0.83 | 0.94 | 1.60 | 0.98 | 0.79 | 1.12 | 1.18 | 3.11 | 2.63 | 1.30 | 2.29 | 2.99 | 1.89 | 1.86 |

Figure A.12: Performance (SSE residual $r^2$) of the variety-specific Seresta model (top) and the variety-agnostic model tested on the Seresta variety (bottom). Rows correspond to the training dataset, columns show the testing configuration. SSE residuals $r^2 \geq 1$ are grayed out.

| Trained on | | 2019 | | | | | | | 2020 | | | | | | | 2021 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Tested on | M | S | V | AVG | MS | MV | SV | M | S | V | AVG | MS | MV | SV | M | S | V | AVG | MS | MV | SV |
| 2019 | M | | | | | 0.68 | | 0.87 | | | | | | | | | | | | | | |
| | S | 0.70 | | | | 0.78 | 0.72 | | | | | | | | | | | | | | | |
| | V | | | | | | 0.83 | | | | | | | | | | | | | | | |
| | AVG | | | | | | | 0.75 | | | | | | | | | | | | | | |
| | MS | 0.79 | | | | | 0.76 | | | | | | | | | | | | | -0.41 | | |
| | MV | | | | 0.80 | 0.72 | | | | | | | | | | | | | -0.38 | -0.45 | | |
| | SV | | | | | | | 0.79 | | | | | | | | | | | | | | |
| 2020 | M | | | | | | | | -0.72 | | | | | | | | | | | -0.38 | | |
| | S | | | | | | | | | | | | 0.80 | 0.82 | | | | | | | | |
| | V | | | | | | | | | | | | | | | | | | -0.40 | -0.41 | | |
| | AVG | | | | | | | | | | | | | | | | | | | -0.37 | | |
| | MS | | | | | | | | 0.74 | | | | | | | | | | | | | |
| | MV | | | | | | | | | 0.86 | | 0.99 | | | | | | | | | | |
| | SV | | | | | | | | | | | 0.70 | | 0.71 | | | | | | | | |
| 2021 | M | | | | | | | | | | | | | | | | | | | | | 0.78 |
| | S | | | | | | | | | | | | | | | | | | | | | |
| | V | | | | | | | | | | | | | | | | | | | | 0.76 | |
| | AVG | | | | | | | | | | | | | | | 0.67 | | | | | 0.84 | |
| | MS | -0.45 | | | | | | | | | | | | | | | | | | | | |
| | MV | -0.55 | | | | -0.48 | | | | | | | | | | | | | | | | |
| | SV | -0.42 | | | | | | | | | | | | | | | | 0.89 | 0.80 | | 0.65 | |

| Trained on | | 2019 | | | | | | | 2020 | | | | | | | 2021 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Tested on | M | S | V | AVG | MS | MV | SV | M | S | V | AVG | MS | MV | SV | M | S | V | AVG | MS | MV | SV |
| 2019 | M | | | | | | | | 0.43 | 0.38 | | | 0.44 | | | | | -0.46 | -0.45 | -0.39 | -0.44 | -0.41 |
| | S | | | | | | | | | | | | | | | | | | | | | |
| | V | | | | | | | | 0.52 | 0.44 | | 0.43 | 0.52 | 0.37 | | | | | | | | |
| | AVG | | | | | | | | 0.66 | 0.62 | | 0.59 | 0.70 | 0.51 | 0.48 | | | | | | | |
| | MS | | | | | 0.79 | | | 0.56 | 0.52 | | 0.49 | 0.58 | 0.41 | 0.39 | | | | | | | |
| | MV | | | | | | | | | | | | | | | | | | | | | |
| | SV | | | | 0.71 | 0.87 | | | 0.51 | 0.40 | | | 0.50 | | | | | | | | | |
| 2020 | M | | | | | | | | | | | | | | | | | | | | | |
| | S | | | | | | | | 0.71 | | | | | | | | | | | | | |
| | V | | | 0.49 | | | | | 0.74 | | | | | | | | | | | | | |
| | AVG | | | | | | | | | | | | | | | | | 0.41 | 0.38 | | 0.40 | |
| | MS | | | | | | | | | | | | | | | 0.41 | | 0.41 | 0.51 | 0.50 | 0.50 | 0.40 |
| | MV | | | | | | | | | | | | | | | | | | | | | |
| | SV | 0.46 | | 0.44 | 0.49 | 0.45 | 0.49 | 0.47 | | | | | | | | | | | | | | |
| 2021 | M | -0.54 | -0.46 | -0.41 | -0.53 | -0.53 | -0.51 | -0.48 | | | | | | | | | | | | | | |
| | S | 0.47 | 0.57 | | | 0.54 | | | | | | | | | | | | | | | | |
| | V | | | | | | | | 0.38 | | | | | | | | | 0.69 | | | | |
| | AVG | | | | | | | | | | | | | | | | | | | | | |
| | MS | -0.55 | | | | -0.42 | | | | | | | | | | | | 0.76 | 0.70 | | | |
| | MV | | | | | | | | 0.43 | 0.37 | | 0.38 | 0.43 | | | | | | | | | |
| | SV | 0.43 | | | | | | | 0.58 | 0.53 | | 0.51 | 0.60 | 0.44 | 0.41 | | | | | | | |

Figure A.13: Performance (Pearson correlation $c$) of the variety-specific Seresta model (top) and the variety-agnostic model tested on the Seresta variety (bottom). Rows correspond to the training dataset, columns show the testing configuration. Correlations with $p \geq 0.05$ are omitted.

# Bibliography

[1] N. Alexandrov, S. Tai, W. Wang, L. Mansueto, K. Palis, R. R. Fuentes, V. J. Ulat, D. Chebotarov, G. Zhang, Z. Li, et al. Snp-seek database of snps derived from 3000 rice genomes. *Nucleic acids research*, 43(D1):D1023–D1027, 2015.

[2] N. An, S. M. Welch, R. C. Markelz, R. L. Baker, C. M. Palmer, J. Ta, J. N. Maloof, and C. Weinig. Quantifying time-series of leaf morphology using 2d and 3d photogrammetry methods for high-throughput plant phenotyping. *Computers and Electronics in Agriculture*, 135:222–232, 2017. ISSN 0168-1699. doi: https://doi.org/10.1016/j.compag.2017.02.001. URL `https://www.sciencedirect.com/science/article/pii/S0168169916302393`.

[3] E. Atza and N. Budko. High-throughput analysis of potato vitality. In M. Ehrhardt and M. Günther, editors, *Progress in Industrial Mathematics at ECMI 2021*, pages 273–279, Cham, 2022. Springer International Publishing.

[4] E. Atza and N. Budko. Overparameterized multiple linear regression as hyper-curve fitting. *SIAM Journal on Mathematics of Data Science (Under review)*, 2024. URL `https://arxiv.org/abs/2404.07849`.

[5] E. Atza and N. Budko. Predicting potato plant vigor from the seed tuber properties. *Scientific Reports (Under review)*, 2024. URL `https://doi.org/10.48550/arXiv.2410.19875`.

[6] E. Atza and N. Budko. Data underlying the publication: Predicting potato plant vigor from the seed tuber properties. *4TU.ResearchData* `http://doi.org/10.4121/3a97fa0c-8c7d-451a-b8fe-d521f1cec55e`, 2024.

[7] E. Atza and N. Budko. Data underlying the publication: Seed tuber microbiome is a predictor of next-season potato vigor. *4TU.ResearchData* `http://doi.org/10.4121/21892a06-078a-4600-8386-1abe46f42271`, 2024.

[8] M. S. Barlett. Nearest neighbour models in the analysis of field experiments. *Journal of the Royal Statistical Society. Series B (Methodological)*, 40(2):147–174, 1978. URL `https://www.jstor.org/stable/2984750`.

[9] A. D. Barnosky, N. Matzke, S. Tomiya, G. O. U. Wogan, B. Swartz, T. B. Quental, C. Marshall, J. L. McGuire, E. L. Lindsey, K. C. Maguire, B. Mersey, and E. A. Ferrer. Has the earth's sixth mass extinction already arrived? *Nature*, 471:51–57, 2011. doi: https://doi.org/10.1038/nature09678.

[10] M. S. Bartlett. Further aspects of the theory of multiple regression. *Mathematical Proceedings of the Cambridge Philosophical Society*, 34(1):33–40, 1938. doi: 10.1017/S0305004100019897.

[11] P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48): 30063–30070, 2020. doi: 10.1073/pnas.1907378117. URL `https://www.pnas.org/doi/abs/10.1073/pnas.1907378117`.

[12] J. Besag and R. Kempton. Statistical analysis of field experiments using neighbouring plots. *Biometrics*, 42(2):231–251, 1986. URL `https://www.jstor.org/stable/2531047`.

[13] H. Boogaard, C. van Diepen, R. Rotter, J. Cabrera, and H. van Laar. *WOFOST 7.1; user's guide for the WOFOST 7.1 crop growth simulation model and WOFOST Control Center 1.5*. Number 52 in Technical document / DLO Winand Staring Centre. Staring Centrum, Netherlands, 1998.

[14] A. Borges, A. González-Reymundez, O. Ernst, M. Cadenazzi, J. Terra, and L. Gutiérrez. Can spatial modeling substitute for experimental design in agricultural experiments? *Crop Science*, 59(1):44–53, 2019. doi: https://doi.org/10.2135/cropsci2018.03.0177. URL `https://acsess.onlinelibrary.wiley.com/doi/abs/10.2135/cropsci2018.03.0177`.

[15] S. Brenner and R. Scott. *The Mathematical Theory of Finite Element Methods*. Springer, 2005.

[16] L. Brigato and L. Iocchi. A close look at deep learning with small data. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 2490–2497, 2021. doi: 10.1109/ICPR48806.2021.9412492.

[17] D. Broadhurst, R. Goodacre, S. N. Reinke, J. Kuligowski, I. D. Wilson, M. R. Lewis, and W. B. Dunn. Guidelines and considerations for the use of system suitability and quality control samples in mass spectrometry assays applied in untargeted clinical metabolomic studies. *Metabolomics*, 14, 2018. doi: 10.1007/s11306-018-1367-3.

[18] S. D. Choudhury, A. Samal, and T. Awada. Leveraging image analysis for high-throughput plant phenotyping. *Frontiers in Plant Science*, 10:436562, 4 2019. ISSN 1664462X. doi: 10.3389/FPLS.2019.00508/BIBTEX.

[19] G. E. M. D. M. Woebbecke and K. V. B. et al. Color indices for weed identification under various soil, residue, and lighting conditions. *Transactions of the ASAE*, 38(1):259–269, 1995. doi: https://doi.org/10.13031/2013.27838.

[20] G. de Los Campos, J. M. Hickey, R. Pong-Wong, H. D. Daetwyler, and M. P. Calus. Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics*, 193(2):327–345, 2013.

[21] K. E. Deshi, M. O. Obasi, and N. I. Odiaka. Growth and yield of potato (solanum tuberosum l.) as affected by storage conditions and storage duration in jos, plateau state, nigeria. *Open Agriculture*, 6:779–797, 1 2021. ISSN 23919531. URL `https://www.degruyter.com/document/doi/10.1515/opag-2021-0057/html`.

[22] D. Di Pietro and A. Ern. *Mathematical Aspects of Discontinuous Galerkin Methods.* Mathématiques et Applications, Vol. 69. Springer-Verlag, Berlin, 2011.

[23] Y. Dong, S. Zhou, L. Xing, Y. Chen, Z. Ren, Y. Dong, and X. Zhang. Deep learning methods may not outperform other machine learning methods on analyzing genomic studies. *Frontiers in Genetics*, 13, Sept. 2022. ISSN 1664-8021. doi: 10.3389/fgene.2022.992070. URL `http://dx.doi.org/10.3389/fgene.2022.992070`.

[24] EC. EU agricultural outlook for markets, 2023-2035. *European Commission and DG Agriculture and Rural Development*, 2024. URL `https://agriculture.ec.europa.eu/data-and-analysis/markets/outlook/medium-term_en`.

[25] R. Escadafal. Remote sensing of arid soil surface color with landsat thematic mapper. *Advances in space research*, 9(1):159–163, 1989.

[26] European Commission. The eu potato sector-statistics on production, prices and trade statistics explained, 2024. URL `https://ec.europa.eu/eurostat/statisticsexplained/`.

[27] Food and Agriculture Organization of the United Nations (FAO) and ESS. Food and Agriculture Organization of the United Nations, *FAOSTAT database*, 2022. URL `https://www.fao.org/faostat/en/#data/QCL`.

[28] Food and Agriculture Organization of the United Nations (FAO), and Statistics Division (ESS). Production / crops and livestock products. `https://www.fao.org/faostat/en/#data/QCL`, 2023.

[29] R. Fritsche-Neto and A. Borém. *Phenomics: How next-generation phenotyping is revolutionizing plant breeding.* Springer International Publishing, 1 2015. ISBN 9783319136776. doi: 10.1007/978-3-319-13677-6/COVER.

[30] A. Gilmour, R. Thompson, and B. Cullis. Average information reml: An efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics*, 51(4):1440–1450, Dec. 1995. ISSN 0006-341X.

[31] A. R. Gilmour, B. R. Cullis, and A. P. Verbyla. Accounting for natural and extraneous variation in the analysis of field experiments. *Journal of Agricultural, Biological, and Environmental Statistics*, 2(3):269–293, 1997. URL https://www.jstor.org/stable/1400446.

[32] T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50 (2):949 – 986, 2022. doi: 10.1214/21-AOS2133. URL https://doi.org/10.1214/21-AOS2133.

[33] T. He, R. Kong, A. J. Holmes, M. Nguyen, M. R. Sabuncu, S. B. Eickhoff, D. Bzdok, J. Feng, and B. T. Yeo. Deep neural networks and kernel regression achieve comparable accuracies for functional connectivity prediction of behavior and demographics. *NeuroImage*, 206:116276, 2020. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2019.116276. URL https://www.sciencedirect.com/science/article/pii/S1053811919308675.

[34] D. Holzmüller. On the universality of the double descent peak in ridgeless regression. *arXiv preprint arXiv:2010.01851*, 2020.

[35] Z. Huang and C. Wang. A review on differential abundance analysis methods for mass spectrometry-based metabolomic data. *Metabolites*, 12:305, 4 2022. ISSN 22181989. doi: 10.3390/METABO12040305. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC9032534/.

[36] S. Huffel and P. Lemmerling. *Total Least Squares and Errors-in-Variables Modeling: Analysis, Algorithms and Applications*. Springer Netherlands, 2002. ISBN 9789401735520. doi: 10.1007/978-94-017-3552-0. URL http://dx.doi.org/10.1007/978-94-017-3552-0.

[37] A. J. Izenman. *Modern Multivariate Statistical Techniques*. Springer New York, 2008. ISBN 9780387781891. doi: 10.1007/978-0-387-78189-1. URL http://dx.doi.org/10.1007/978-0-387-78189-1.

[38] S. Janssens, S. Wiersema, H. Goos, and W. wiersma. The value chain for seed and ware potatoes in kenya: Opportunities for development, 2013. URL www.wageningenUR.nl/en/lei.

[39] E. Kolbert. *The Sixth Extinction: An Unnatural History*. Henry Holt and Company, 2014. ISBN 978-0-8050-9299-8.

[40] T. K. Kwambai, D. Griffin, M. Nyongesa, S. Byrne, M. Gorman, Paul, and C. Struik. Dormancy and physiological age of seed tubers from a diverse set of potato cultivars grown at different altitudes and in different seasons in kenya. *Potato Research*, 2023. doi: 10.1007/s11540-023-09617-9. URL https://doi.org/10.1007/s11540-023-09617-9.

[41] S. H. Lee and J. H. Van der Werf. Mtg2: an efficient algorithm for multivariate linear mixed model analysis based on genomic information. *Bioinformatics*, 32 (9):1420–1422, 2016.

[42] Z. Li, R. Guo, M. Li, Y. Chen, and G. Li. A review of computer vision technologies for plant phenotyping. *Computers and Electronics in Agriculture*, 176:105672, 2020. ISSN 0168-1699. doi: https://doi.org/10.1016/j.compag.2020.105672. URL `https://www.sciencedirect.com/science/article/pii/S0168169920307511`.

[43] R. Linker, I. Shmulevich, A. Kenny, and A. Shaviv. Soil identification and chemometrics for direct determination of nitrate in soils using FTIR-ATR mid-infrared spectroscopy. *Chemosphere*, 61(5):652–658, 2005. ISSN 0045-6535. doi: https://doi.org/10.1016/j.chemosphere.2005.03.034. URL `https://www.sciencedirect.com/science/article/pii/S0045653505004455`.

[44] A. Lorence and K. M. Jimenez, editors. *High-Throughput Plant Phenotyping*, volume 2539. Springer US, 2022. ISBN 978-1-0716-2536-1. doi: 10.1007/978-1-0716-2537-8. URL `https://link.springer.com/10.1007/978-1-0716-2537-8`.

[45] R. Mathieu, M. Pouget, B. Cervelle, and R. Escadafal. Relationships between satellite-based radiometric indices simulated using laboratory reflectance data and typic soil color of an arid environment. *Remote Sensing of Environment*, 66(1):17–28, 1998. ISSN 0034-4257. doi: https://doi.org/10.1016/S0034-4257(98)00030-3. URL `https://www.sciencedirect.com/science/article/pii/S0034425798000303`.

[46] L. Messeri and M. Crockett. Artificial inteligence and illusions of understanding in scientific research. *Nature*, 4627(8002):49–58, 2024.

[47] B.-H. Mevik and R. Wehrens. The pls package: Principal component and partial least squares regression in r. *Journal of Statistical Software*, 18(2):1–23, 2007. doi: 10.18637/jss.v018.i02. URL `https://www.jstatsoft.org/index.php/jss/article/view/v018i02`.

[48] Nederlandse Algemene Keuringsdienst, 2022. URL `https://www.nak.nl/publicaties/inspection-of-seed-potatoes/`.

[49] M. Oh and L. Zhang. Deepmicro: deep representation learning for disease prediction based on microbiome data. *Scientific Reports*, 10(1), Apr. 2020. ISSN 2045-2322. doi: 10.1038/s41598-020-63159-5. URL `http://dx.doi.org/10.1038/s41598-020-63159-5`.

[50] J. S. Papadakis. Methode statistique pour des experiences sur champ. *Bull. Inst. Amel. Plantes a' Salonique*, 23, 1937.

[51] R. Raymundo, S. Asseng, D. Cammarano, and R. Quiroz. Potato, sweet potato, and yam models for climate change: A review. *Field Crops Research*, 166:173–185, 9 2014. ISSN 0378-4290. doi: 10.1016/J.FCR.2014.06.017.

[52] M. X. Rodríguez-Álvarez, M. P. Boer, F. A. van Eeuwijk, and P. H. C. Eilers. Correcting for spatial heterogeneity in plant breeding experiments with p-splines. *Spatial Statistics*, 23:52–71, 2018. ISSN 2211-6753. doi: https://doi.org/10.1016/j.spasta.2017.10.003. URL `https://www.sciencedirect.com/science/article/pii/S2211675317301070`.

[53] M. X. Rodríguez-Álvarez, M. P. Boer, F. A. van Eeuwijk, and P. H. C. Eilers. https://cran.r-project.org/package=spats, 2018.

[54] N. L. Rozali, K. A. Azizan, R. Singh, S. N. Syed Jaafar, A. Othman, W. Weckwerth, and U. S. Ramli. Fourier transform infrared (FTIR) spectroscopy approach combined with discriminant analysis and prediction model for crude palm oil authentication of different geographical and temporal origins. *Food Control*, 146:109509, 2023. ISSN 0956-7135. doi: https://doi.org/10.1016/j.foodcont.2022.109509. URL `https://www.sciencedirect.com/science/article/pii/S0956713522007022`.

[55] C. J. Sands, A. M. Wolfer, G. D. S. Correia, N. Sadawi, A. Ahmed, B. Jiménez, M. R. Lewis, R. C. Glen, J. K. Nicholson, and J. T. M. Pearce. The nPYc-Toolbox, a Python module for the pre-processing, quality-control and analysis of metabolic profiling datasets. *Bioinformatics*, 35(24):5359–5360, 07 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz566. URL `https://doi.org/10.1093/bioinformatics/btz566`.

[56] S. Sehrawat, K. Najafian, and L. Jin. Predicting phenotypes from novel genomic markers using deep learning. *Bioinformatics Advances*, 3(1):vbad028, 03 2023. ISSN 2635-0041. doi: 10.1093/bioadv/vbad028. URL `https://doi.org/10.1093/bioadv/vbad028`.

[57] M. Sheikh, F. Iqra, H. Ambreen, K. A. Pravin, M. Ikra, and Y. S. Chung. Integrating artificial intelligence and high-throughput phenotyping for crop improvement. *Journal of Integrative Agriculture*, 23(6):1787–1802, 2024. ISSN 2095-3119. doi: https://doi.org/10.1016/j.jia.2023.10.019. URL `https://www.sciencedirect.com/science/article/pii/S2095311923003611`.

[58] Y. Song, E. Atza, J. J. Sanchez Gil, D. Akkermans, R. de Jonge, P. G. de Rooij, D. Kakembo, P. A. Bakker, C. M. Pieterse, N. V. Budko, and R. L. Berendsen. Seed tuber microbiome can predict growth potential of potato varieties. *Nature Microbiology*, 10:28–40, 2025. ISSN 2058-5276. doi: 10.1038/s41564-024-01872-x. URL `https://doi.org/10.1038/s41564-024-01872-x`.

[59] M. Steinfath, N. Strehmel, R. Peters, N. Schauer, D. Groth, J. Hummel, M. Steup, J. Selbig, J. Kopka, P. Geigenberger, et al. Discovering plant

metabolic biomarkers for phenotype prediction using an untargeted approach. *Plant Biotechnology Journal*, 8(8):900–911, 2010.

[60] H. Swierenga, A. de Weijer, R. van Wijk, and L. Buydens. Strategy for constructing robust multivariate calibration models. *Chemometrics and Intelligent Laboratory Systems*, 49(1):1–17, 1999. ISSN 0169-7439. doi: https://doi.org/10.1016/S0169-7439(99)00028-3. URL `https://www.sciencedirect.com/science/article/pii/S0169743999000283`.

[61] E. V. Thomas. A primer on multivariate calibration. *Analytical Chemistry*, 66 (15):795A–804A, 1994.

[62] C. van Diepen, J. Wolf, H. van Keulen, and C. Rappoldt. Wofost: a simulation model of crop production. *Soil Use and Management*, 5 (1):16–24, 1989. doi: https://doi.org/10.1111/j.1475-2743.1989.tb00755.x. URL `https://bsssjournals.onlinelibrary.wiley.com/doi/abs/10.1111/j.1475-2743.1989.tb00755.x`.

[63] C. R. Vogel. *Computational methods for inverse problems.* Frontiers in applied mathematics. SIAM, 2002. ISBN -89871-507-5.

[64] K. Vuik, F. Vermolen, M. van Gijzen, and T. Vuik. *Numerical Methods for Ordinary Differential Equations.* VSSD, TU Delft Open, 2023. doi: https://doi.org/10.5074/t.2023.001.

[65] K. Vuik, F. Vermolen, M. van Gijzen, and T. Vuik. Numerical methods for ordinary differential equations. *TU Delft OPEN Textbooks*, 2 2023. doi: 10.5074/T.2023.001.

[66] J. Wei, A. Wang, R. Li, H. Qu, and Z. Jia. Metabolome-wide association studies for agronomic traits of rice. *Heredity*, 120(4):342– 355, 2018. doi: https://doi.org/10.1038/s41437-017-0032-3. URL `https://www.nature.com/articles/s41437-017-0032-3#Sec1`.

[67] B. T. West, K. B. Welch, and A. T. Galecki. *Linear mixed models: a practical guide using statistical software.* Chapman and Hall/CRC, 2022.

[68] C. C. Westhues, G. S. Mahone, S. da Silva, P. Thorwarth, M. Schmidt, J.-C. Richter, H. Simianer, and T. M. Beissinger. Prediction of maize phenotypic traits with genomic and environmental predictors using gradient boosting frameworks. *Frontiers in plant science*, 12:699589, 2021.

[69] C. Wilson. An errors-in-variables spatial mixed model. *unpublished M. Lit. Thesis, University of New England, Armidale, NSW*, 1994.

[70] S. Wold, M. Sjöström, and L. Eriksson. PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58(2):109–130, 2001. ISSN 0169-7439. doi: https://doi.org/10.1016/S0169-7439(01)

00155-1. URL https://www.sciencedirect.com/science/article/pii/S0169743901001551. PLS Methods.

[71] K. Zarzyńska, D. Boguszewska-Mańkowska, B. Feledyn-Szewczyk, and K. Jończyk. The vigor of seed potatoes from organic and conventional systems. *Agriculture (Switzerland)*, 12, 11 2022. ISSN 20770472. doi: 10.3390/agriculture12111764.

[72] D. L. Zimmerman and D. A. Harville. A random field approach to the analysis of field-plot experiments and other spatial experiments. *Biometrics*, 47(1):223–239, 1991. URL https://www.jstor.org/stable/2532508.

# Summary

Potato plant vitality is an important trait for good yield. Unfortunately, even with the best potato seed, vitality variation within a genotype is significant and it hurts both farmers and potato seed producers. HZPC and Averis Seeds hypothesize the existence of a link between the variation in the vitality of a plant and the field of production of the seed tuber. TU Delft, Utrecht University, HZPC and Averis Seeds joined forces to determine the existence and strength of this link in a multi-year project funding this research. In this thesis we describe TU Delft's contribution both in the measurement of plant vitality and the development of a model to predict plant vigor from chemical properties of the tuber.

We start this thesis with a discussion of the results achieved and of the goals that were not attained. We give some context as well as some suggestions so that future projects can learn from our experience.

Our work in the first years focused on the extraction of a precise vitality measurement from field and climate controlled room imaging. To this purpose we developed software to deal with specific challenges connected to our project setup and goals. In Chapter 2 we describe the experimental design both in the fields and in the climate rooms, we outline the main steps leading from a field image to a measurement of vitality and describe in detail the image processing and the choices made to arrive at a single vitality measure per field. From the vitality measurement is already possible to confirm the project's hypothesis by observing a strong, statistically significant Pearson correlation coefficient between canopy areas in different fields.

The development of a predictive model for vitality has been an iterative process, starting from the chemometric data (FTIR, HSI, and XRF) and gradually expanding to encompass the microbiome and metabolome. Our initial association analysis used Partial Least Squares (PLS) to connect vitality to the chemometric data. This is a method developed by Wold in 1974 to solve linear problems with high collinearity in the independent variables, such as those arising in chemometrics applications. As we expanded our model to include biotic data, we noticed that linear methods failed to predict. We developed the PARCUR method, described in Chapter 3 which inverts the usual regression approach by expressing the columns of the matrix of features

$X$ as functions of the dependent variable $\boldsymbol{y}$. This approach is customizable in that the representation basis can be adapted to model specific needs, the choice of basis can also carry a built in regularization approach and lastly the method measures the quality of feature representation per feature, this is used to discard uninformative features without need to retrain. In this chapter we illustrate the method both on a synthetic dataset and on a publicly available FTIR dataset for PET-yarn.

Finally in Chapter 4 we apply the "PARCUR" method to the data of three years. We combine chemical and metabolic data in this model, for this purpose we use the modified version of "PARCUR" which represents features in a Discontinuous Galerkin basis. We confirm some of the conclusions reached by the microbiome predictive model, i.e. the degree of precision to which we can predict vitality varies with the variety. We also restrict the set of predictive features considerably, attaining precise results with one of the cheapest-to-measure datasets: FTIR.

In conclusion our model can predict the vitality of a genotype from the seed tuber composition, but within the genotype the model is only capable of predicting a seedlot's vitality for some of the studied genotypes, generally, those genotypes whose vitality is not too strong. It is our belief that given the strong influence of environmental variables on the field data, the sample size in this project was too small to achieve a meaningful representation of the true seedlot distribution.

# Samenvatting

De vitaliteit van aardappelplanten is een belangrijk kenmerk voor een goede opbrengst. Helaas is zelfs met het beste aardappelzaad de vitaliteitsvariatie binnen een genotype aanzienlijk en dit schaadt zowel boeren als aardappelzaadproducenten. HZPC en Averis Seeds stellen de hypothese dat er een verband bestaat tussen de variatie in vitaliteit van een plant en het productieveld van de pootknol. TU Delft, Universiteit Utrecht, HZPC en Averis Seeds hebben hun krachten gebundeld om het bestaan en de sterkte van dit verband vast te stellen in een meerjarig project dat dit onderzoek financiert. In dit proefschrift beschrijven we de bijdrage van de TU Delft, zowel in het meten van plantvitaliteit als de ontwikkeling van een model om plantvitaliteit te voorspellen op basis van chemische eigenschappen van de knol.

We beginnen dit proefschrift met een bespreking van de bereikte resultaten en van de doelen die niet zijn bereikt. We geven wat context en enkele suggesties zodat toekomstige projecten kunnen leren van onze ervaring.

Ons werk in de eerste jaren richtte zich op het extraheren van een nauwkeurige vitaliteitsmeting uit veld- en klimaatgecontroleerde kamerbeelden. Voor dit doel ontwikkelden we software voor specifieke uitdagingen die samenhingen met onze projectopzet en -doelen. In hoofdstuk 2 beschrijven we de proefopzet, zowel op het veld als in de klimaatkamers, schetsen we de belangrijkste stappen die leiden van een veldopname tot een meting van vitaliteit en beschrijven we in detail de beeldverwerking en de keuzes die gemaakt zijn om tot een enkele vitaliteitsmeting per veld te komen. Op basis van de vitaliteitsmeting is het al mogelijk om de hypothese van het project te bevestigen door een sterke, statistisch significante Pearson correlatiecoëfficiënt tussen de kroonoppervlakten in verschillende velden te observeren.

De ontwikkeling van een voorspellend model voor vitaliteit is een iteratief proces geweest, beginnend bij de chemometrische data (FTIR, HSI en XRF) en geleidelijk aan uitgebreid met het microbioom en metaboloom. Onze eerste associatieanalyse gebruikte Partial Least Squares (PLS) om vitaliteit te verbinden met de chemometrische data. Dit is een methode die in 1974 door Wold werd ontwikkeld om lineaire problemen op te lossen met een hoge collineariteit in de onafhankelijke variabelen, zoals die zich voordoen in chemometrische toepassingen. Toen we ons model uitbreidden met biotische data, merkten we dat lineaire methoden er niet in slaagden te voorspellen. We ontwikkelden de PARCUR-methode, beschreven in hoofdstuk 3, die de gebruikelijke regressiebenadering omkeert door de kolommen van de matrix van kenmerken X uit te drukken als functies van de afhankelijke variabele y. Deze benadering is aanpasbaar in die zin dat de representatiebasis kan worden aangepast aan modelspecifieke behoeften, de keuze van de basis ook een ingebouwde regularisatieaanpak kan bevatten en tot slot de methode de kwaliteit van de representatie van kenmerken per kenmerk meet, wat wordt gebruikt om niet-informatieve kenmerken te verwijderen zonder dat opnieuw hoeft te worden getraind. In dit hoofdstuk illustreren we de methode zowel op een synthetische dataset als op een publiek beschikbare FTIR dataset voor PET-garen.

Tot slot passen we in hoofdstuk 4 de “PARCUR”-methode toe op de data van drie jaar. In dit model combineren we chemische en metabole data. Hiervoor gebruiken we de gewijzigde versie van “PARCUR”, die kenmerken weergeeft in een Discontinuous Galerkin basis. We bevestigen een aantal conclusies van het microbioom voorspellende model, namelijk dat de mate van precisie waarmee we vitaliteit kunnen voorspellen varieert met de soort. We beperken ook de verzameling van voorspellende kenmerken aanzienlijk en bereiken nauwkeurige resultaten met een van de goedkoopste datasets om te meten: FTIR.

Concluderend kan ons model de vitaliteit van een genotype voorspellen op basis van de pootknolsamenstelling, maar binnen het genotype is het model alleen in staat om de vitaliteit van een partij pootgoed te voorspellen voor sommige van de bestudeerde genotypen, over het algemeen die genotypen waarvan de vitaliteit niet te sterk is. We zijn ervan overtuigd dat gezien de sterke invloed van omgevingsvariabelen op de velddata, de steekproefgrootte in dit project te klein was om een zinvolle weergave van de werkelijke pootgoedverdeling te bereiken.

# Acknowledgements

I would like to thank my daily supervisor, Neil, for his guidance, support, and collaboration, for the time spent talking maths and life and the care he puts in his job.

A big thanks goes to Kees for his guidance and knowledge in the research and the many processes and bureaucracy behind it.

This research has been funded by HZPC B.V. and Averis Seeds. I would like to thank everyone involved in the project Flight to Vitality and particularly Falko Hofstra, Hans van Doorn, Johan Hopman, Rob Klooster, and Frank van der Werff.

I would like to thank the plant-microbiome interaction group of Utrecht University for the nice exchange and fruitful collaboration, particularly Yang Song and Roeland Berendsen for the many discussions which taught me a lot about interdisciplinary communication.

Over the years at TU Delft and the many offices I occupied, many people have contributed to an enjoyable working atmosphere.

In the early days surely the probability and statistics groups that adopted me in building 28, the Numerical Analysis PhD group, the unforgettable first office with Bart, Mario, Rik, the cursed desk and the planking.

Thanks to Richard and Jasper for the daily lunch and discussion sessions in the summer of 2020 that gave a semblance of normalcy to a very weird time.

Thanks to Merel, a great office mate who became a friend with amazing podcast and music recommendations.

I would like to thank the colleagues from the board of the PhD Forum: Vicente, Naqi, Frederik, Matthijs, and Lasse, for the good times organizing events for reluctant phds.

Thanks to Bart and Andrea, your support and the restaurant research in Rotterdam played a big role in my sanity.

A big thanks goes to the Undercover Book Club for the PhD support and for being an amazing group of people, particularly thanks to Carolina, Gina, and Laurane for the spin-off hikes.

Furthermore I would like to thank the long-time long-distance friends: Giulia, Carlotta, Veronica, Silvia, Antonella, Raffaella, Esmee, Peggy, Albert, Nestor, and Łukasz which have been part of these challenging years and the challenging years before. Danke an Jessy für die alten Zeiten und die letzten Monate.

Danke an meine deutsche Familie, die mich immer willkommen geheißen hat und mir in all den Jahren ein Umfeld zum Entspannen, Arbeiten oder für Quarantäneaufenthalte geboten hat.

Grazie alla mia famiglia per avermi sempre supportata nelle scelte più disparate, e particolarmente alle mie sorelle, che hanno deciso di trasferirsi sulla linea del Thalys dando inizio a una tradizione di visite e abbuffate.

To Felix, my number one fan, for having been by my side for a decade with unwavering love and optimism, and for always believing in me more than anyone else.

# Curriculum Vitae

Elisa Atza was born in Sardinia where she graduated from the Liceo Classico S.A. De Castro in Oristano. She obtained a Bachelor in Mathematics at the University of Bologna and then moved to Bonn, Germany where she completed a Master of Science in Mathematics at the University of Bonn with a master thesis on deformations of Lie Groupoids under the supervision of Dr. Christian Blohmann of the Max Planck Institut für Mathematik. After her Master studies she worked as a consultant for d-fine GmbH for different clients all over Germany. This experience motivated her to pursue a PhD in Mathematics. For this she moved to the Netherlands in order to work with Dr. Neil Budko on the project Flight to Vitality.

# Publications

## Published

[3] E. Atza and N. Budko. High-throughput analysis of potato vitality. In M. Ehrhardt and M. Günther, editors, *Progress in Industrial Mathematics at ECMI 2021*, pages 273–279, Cham, 2022. Springer International Publishing

[58] Y. Song, E. Atza, J. J. Sanchez Gil, D. Akkermans, R. de Jonge, P. G. de Rooij, D. Kakembo, P. A. Bakker, C. M. Pieterse, N. V. Budko, and R. L. Berendsen. Seed tuber microbiome can predict growth potential of potato varieties. *Nature Microbiology*, 10:28–40, 2025. ISSN 2058-5276. doi: 10.1038/s41564-024-01872-x. URL `https://doi.org/10.1038/s41564-024-01872-x`

## Submitted

[4] E. Atza and N. Budko. Overparameterized multiple linear regression as hyper-curve fitting. *SIAM Journal on Mathematics of Data Science (Under review)*, 2024. URL `https://arxiv.org/abs/2404.07849`

[5] E. Atza and N. Budko. Predicting potato plant vigor from the seed tuber properties. *Scientific Reports (Under review)*, 2024. URL `https://doi.org/10.48550/arXiv.2410.19875`