High-Throughput Analysis of Potato Vitality

Elisa Atza and Neil Budko

Abstract Vitality is a fundamental trait for the development of a plant. It is known to depend on various factors, such as climate, soil, and the plant's genetics, but the progressive depletion of soil nutrients make it a priority for the industry to pinpoint which of the controllable qualities of a seed have the biggest impact on vitality. This work describes techniques applied in a high-throughput phenotyping project, the first of this magnitude for a complex plant, the potato (*solanum tuberosum*). We also present the results of an analysis of associations between the chemical composition of the seed potatoes and field performance, solving the arising underdetermined linear systems by means of PLS regression. We show that some but not all of the chemical data is strongly associated to vitality.

1 Introduction

A potato plant is vital if it manifests in a large canopy and exhibits homogeneous growth in the early stages of its development. Potato seed producers as well as farmers have noticed that potato seeds of the same cultivar perform differently in the same conditions depending on the field in which the seed tubers have been produced.

A cultivar, or variety, is described as a set of plants for which specific characteristics are reliably passed on to the offspring. Uniform growth facilitates farming thus high variability in the growth of a variety is undesirable.

In collaboration with potato seed producers HZPC and Averis seeds, we aim to quantify the contribution of non genetic factors to the plant development. Specifically, we investigate the link between the chemical and biological properties of a tuber, and the vitality of the sprouting plant. Identifying relevant markers would allow for a screening of tubers prior to planting; allowing for higher yields and customised offers to the clients. There is no standard way to relate such a broad

E. Atza (✉) · N. Budko
Delft Institute of Applied Mathematics, Delft University of Technology, Delft, The Netherlands
e-mail: e.atza@tudelft.nl; n.v.budko@tudelft.nl

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_36

variety of interdependent data regarding a tuber to the measured development of the plant.

In order to model this problem we study six different cultivars, and for each cultivar we measure 30 different tubers, which are genetically identical, but have either been produced in different locations or have received a specific treatment. These 30 different tubers we call batches, so that in total we study 180 different batches belonging to 6 varieties.

For the experiment, data was collected from the studied tubers before and after planting over several consecutive years. Our industrial partner, HZPC, is responsible for the collection of most of the tuber data, as well as for the planting process. Aerial pictures are collected for the field experiment in different European locations by a commercial drone operator, which provides us with orthophotos of the fields according to industrial standard. We then process these aerial pictures ourselves in order to quantify vitality from expressed traits of the plant, a process referred to as phenotyping.

We will shortly present the procedure used to extract canopy coverage in the field from drone images, and then discuss the first associations resulting from linear regression performed considering the different data sets as independent variables.

2 Linear Regression with PLS

We predict vitality parameters $Y \in \mathbb{R}^{180}$, from different tuber data $X \in \mathbb{R}^{180\times p}$, by investigating the presence of a linear dependence:

$$Y = X\beta + \epsilon. \tag{1}$$

2.1 Response

The experiment fields are planted according to a randomized block design, so that four replicates of the 180 batches are distributed on separate non-adjacent parts of the field. The first step in the processing of aerial pictures is to delimit the regions, called plots, where each batch is planted.

For each field we choose one image dated around 35–40 days after planting to find these boundaries. At this point in their development, plants within one plot form a continuous canopy and simultaneously have not grown enough to bridge the gap to the next plot. Thus, looking for gaps in the vegetation at this stage almost coincides with looking for plot boundaries.

We use both physical markers on the field and manual input to determine the region of interest in the drone image, then algorithmically look for gaps inside this region. Knowing the number of plots and the number of columns (*ridges*) in each portion of the field, our algorithm determines the most likely plot boundaries.

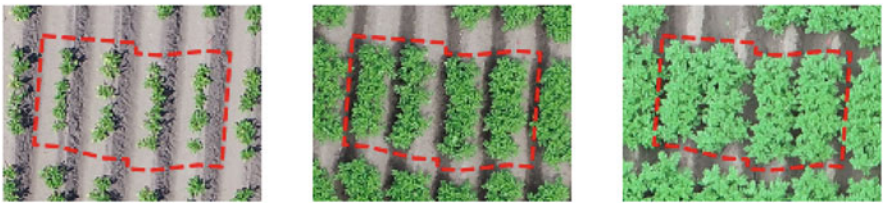

Fig. 1 Plot boundaries are detected in the middle image, these boundaries are then used on other photos after time alignment. Each plot is subdivided in four columns, called ridges

After having determined and visually inspected the plot boundaries found on this date, we use physical marks present on the field to align all photos of the same field, such that the boundaries found can be used both before and after the reference date, when canopies' growth makes it harder to distinguish plots, or when dealing with delayed sprouting and small canopies. An example is given in Fig. 1.

Given the resolution of the orthophotos, we look at four vitality measurements per plot, namely we quantify the mean canopy coverage in each ridge. In this way we obtain 16 canopy measurements (four ridges times four plots) for each of the 180 batches on any of the r measurement dates, i.e. our response $\mathbf{Y} \in \mathbb{R}^{2880 \times r}$.

The mean ridge canopy data must be corrected for possible smooth spatial variations across the field due to large-scale inhomogeneities in soil properties and other factors influencing the growth of plants.

For each measurement date $j = 1, \ldots, r$ we model the spatial variations in each column $\mathbf{Y}(j)$ as

$$\mathbf{Y}(j) = X_1 \beta_1 + \boldsymbol{\epsilon}_1, \quad \boldsymbol{\epsilon}_1 \sim \mathcal{N}(0, \sigma^2 I), \tag{2}$$

where $\beta_1 = [c_1, c_2, c_3, c_4]^T \in \mathbb{R}^{p_1}$, $p_1 = 4$. The structure of the design matrix $X_1 \in \mathbb{R}^{n \times p_1}$, $n = 2880$, can be inferred from (3), which is the row-wise expression of (2), and σ^2 is the field specific variance, which we estimate from the data.

For a single ridge i, $i = 1, \ldots, 2880$, located at pixel coordinates $\langle x_i, y_i \rangle$ the model in (2) translates to:

$$Y_i = c_1 + c_2 x_i + c_3 y_i + c_4 x_i y_i + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2), \tag{3}$$

For spatial correction we retain the field mean, but the global linear and bi-linear spatial variations are removed:

$$\mathbf{Y}_{\text{corr}}(j) = \mathbf{Y}(j) - X_1 \hat{\beta}_1 + \hat{c}_1 \mathbf{1} = \hat{\boldsymbol{\epsilon}}_1 + \hat{c}_1 \mathbf{1}. \tag{4}$$

where $\hat{\beta}_1$ is the restricted maximum likelihood (REML) estimate of β_1 and $\hat{\boldsymbol{\epsilon}}_1 \sim \mathcal{N}(0, \hat{\sigma}^2 I)$, where $\hat{\sigma}^2$ is the REML estimate of σ^2.

After correction we consider the average growth performance of a batch over multiple days and multiple repetitions reducing the size of our response to $Y \in$

$\mathbb{R}^{180\times 1}$. This is then normalized to have zero mean and unit standard deviation and is our response for the model in (1).

2.2 Predictors

Several aspects of the tubers are analyzed in the scope of the project with the goal of obtaining an exhaustive description of the chemical and biological profile of different batches of the same variety.

In this work we look at three tuber related datasets and we will compare their performance as predictors of vitality:

- Fourier transform infrared (**FTIR**) spectroscopy: for each sample we obtain a spectrum, i.e. a discretized curve, whose values are the absorbances of the sample for given wavenumbers, in this case the matrix X is of size 180×2388.
- Hyperspectral imaging (**HSI**): each sample is photographed at l different wavelenghts resulting in l images of size $w \times h$. This results in an array of size $w \times h \times l$. The values of a pixel at different wavelengths form an array of length l. Averaging these arrays over particular regions of the tuber we obtain spectra for known tuber compartments, such as pith and cortex. In this case $l = 288$, thus we obtain for each compartment a matrix of predictors X of size 180×288.
- X-Ray fluorescence (**XRF**): this technique gives us concentrations of 10 chemical elements in the samples, in this case the predictor matrix X has size 180×10.

Also for our predictors we apply a zero mean and unit standard deviation normalization. Additionally, for the spectral data (FTIR, HSI) we explore normalization by applying the Savitzky-Golay first polynomial derivative (SG1).

For two data sets (FTIR, and HSI) the linear model in (1) is highly underdetermined. We use partial least squares (PLS) regression to solve the resulting system of equations.

2.3 Method

PLS, also called projection to latent structures, is a dimensionality reduction technique for which the explanatory and the dependent variables are both projected on new *components* constructed to maximize the covariance between X and Y, see [1] and [2]. The decomposition of both matrices X and Y is given by the following:

$$X = T P^T + E, \quad T = (\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_k), \tag{5}$$

$$Y = U Q^T + F, \quad U = (\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k). \tag{6}$$

Here, T, U contain the k latent vectors as columns. Matrices P, Q, are the matrices of loadings, E and F are the residuals.

The columns of the matrices T and U, the latent vectors, are constructed iteratively by finding weights w_i and c_i for which $t_i = Xw_i$ and $u_i = Yc_i$ with the constraints that $w_i^T w_i = 1$, $t_i^T t_i = 1$ and such that $t_i^T u_i$, which is proportional to the covariance of t_i and u_i, is maximal. Each subsequent column is constructed to be orthogonal to the previous ones, and lastly the matrix T of latent vectors for X is used to predict Y with ordinary least squares (OLS). The maximum number of components of the matrix T is equal to the rank of X, at which point the PLS estimator for the coefficients β will be equal to the minimum length least square estimator, [3], which will have large variance for highly collinear spectroscopic data, [4], thus choosing the number of components to be used is a critical point in the application of this method.

As is usual we split our data in train and test set, in order to find the appropriate number of PLS components, we train models with an increasing number of components up to a preset maximum and at each iteration we use k-fold stratified cross validation, $k = 10$, to evaluate the mean squared error, MSE. We choose then the number of components for which the mean of the k MSEs was minimal. Then we train a model with the optimal number of components, which we evaluate on the test set using the coefficient of determination, R^2, and MSE.

This splitting and training is repeated multiple times, the scores R^2 and MSE are stored in the vectors $\mathbf{R^2}$, and **MSE** respectively, so that we can test the robustness of our model by making sure that the empirical standard deviations of the vectors $\sigma(\mathbf{R^2})$ and $\sigma(\mathbf{MSE})$ have a sufficiently small value.

For the XRF dataset we estimate the regression coefficients with OLS.

3 Results

All data in Tables 1, 2, 3, and 4 is displayed in ascending order with respect to the mean R^2. The regression scores are presented for each field and for each year separately. In the case of spectroscopic data we present the results obtained for different normalizations of the tuber data on separate lines.

From our analysis we notice a strong association of the FTIR data set to vitality, regardless of the applied normalization, our evaluation parameters stay consistent for each field.

The regression on FTIR and HSI spectra shows that the prediction performance is influenced by the field in which the vitality has been measured. Furthermore we see that XRF as a stand-alone dataset is not a sufficiently good predictor of vitality, and that the subdivision of hyperspectral data in separate tuber compartments does not offer a substantial difference in performance.

Table 1 Regression on HSP data for year 2

Field	Part	Normal.	# comp	$\mu(R^2)$	$\sigma(R^2)$	μ(MSE)	σ(MSE)
C	pith	STD	38	0.27	0.05	0.27	0.02
C	pith	SG1	39	0.27	0.05	0.27	0.02
C	cortex	SG1	39	0.30	0.04	0.26	0.01
C	cortex	STD	39	0.30	0.04	0.26	0.01
B	pith	STD	33	0.38	0.06	0.39	0.04
B	pith	SG1	37	0.38	0.06	0.39	0.04
B	cortex	SG1	39	0.38	0.06	0.38	0.04
B	cortex	STD	39	0.38	0.06	0.38	0.04
A	pith	SG1	37	0.44	0.08	0.40	0.06
A	pith	STD	38	0.44	0.08	0.40	0.05
A	cortex	STD	39	0.48	0.06	0.38	0.05
A	cortex	SG1	39	0.48	0.06	0.38	0.05

Table 2 Regression on XRF data for both years

Field	Year	$\mu(R^2)$	$\sigma(R^2)$	μ(MSE)	σ(MSE)
C	2	0.21	0.03	0.32	0.04
B	2	0.24	0.05	0.37	0.03
C	1	0.33	0.04	0.36	0.01
A	2	0.33	0.04	0.32	0.02
B	1	0.42	0.02	0.39	0.03
A	1	0.44	0.01	0.31	0.02

Table 3 Regression on FTIR data for year 1

Field	Normal.	# comp.	$\mu(R^2)$	$\sigma(R^2)$	μ(MSE)	σ(MSE)
C	STD	28	0.67	0.02	0.17	0.01
C	SG1	30	0.67	0.02	0.17	0.01
B	STD	19	0.81	0.01	0.12	0.01
B	SG1	21	0.81	0.01	0.12	0.01
A	STD	30	0.84	0.02	0.09	0.01
A	SG1	29	0.84	0.02	0.09	0.01

Table 4 Regression on FTIR data for year 2

Field	Normal.	# comp.	$\mu(R^2)$	$\sigma(R^2)$	μ(MSE)	σ(MSE)
C	SG1	14	0.62	0.02	0.14	0.01
C	STD	14	0.62	0.03	0.14	0.01
B	SG1	22	0.80	0.01	0.13	0.01
B	STD	22	0.80	0.01	0.13	0.01
A	STD	23	0.84	0.02	0.11	0.01
A	SG1	24	0.84	0.02	0.11	0.01

4 Conclusions and Further Research

FTIR data is the best performing, and most consistent in predictive power over different years. Ongoing research suggests that a more tailored analysis of HSI data could improve its predictive performance. Furthermore, the strong link between prediction performance and field of measurement, as well as the fitness of non-linear models for regression on chemical datasets should be investigated.

References

1. P.H. Garthwaite, An Interpretation of Partial Least Squares, Journal of the American Statistical Association 89, No. 425 (1994), 122–27,
2. M. Haenlein and A.M. Kaplan, A Beginner's Guide to Partial Least Squares Analysis, Understanding Statistics, 3:4 (2004), 283–297.
3. A. Phatak and F. de Hoog, Exploiting the connection between PLS, Lanczos methods and conjugate gradients: alternative proofs of some properties of PLS, J. Chemometrics, 16 (2002), 361–367.
4. S. Wold, M. Sjöström, and L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemometrics and Intelligent Laboratory Systems, Volume 58, Issue 2, 2001, Pages 109–130,