KU LEUVEN

FACULTY OF SCIENCE

Comparison of statistical methods and designs for a high throughput phenotyping experiment

Alexandre BOHYN

Supervisor: Prof. P. Goos
KULeuven

Mentor: Pr. X. Draye
UCLouvain

Thesis presented in
fulfillment of the requirements
for the degree of Master of Science
in Statistics

Academic year 2018-2019

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Chapter 2

Literature review

2.1 Plant phenotyping

The terms phenotype and phenotyping are often interpreted in diverse ways between authors and between studies. In order to avoid any confusion, it is important to define these concepts clearly[1]. plant phenotyping is defined as the identification of effects on the phenotype (i.e., the plant appearance and performance) as a result of genotype differences (i.e., differences in the genetic code) and the environmental conditions to which a plant has been exposed (Fiorani & Schurr 2013; Houle *et al.* 2010). In this thesis, we refer to phenotyping more precisely as the set of methodologies and protocols used to measure plant growth, architecture, and composition with a certain accuracy and precision at different scales of organization (Fiorani & Schurr 2013).
Plant phenotyping is an important tool to address and understand plant environment interaction and its translation into application in crop management practices, effects of biostimulants, microbial communities, etc... (Pieruschka, Schurr, *et al.* 2019). In our current society, food security is a rising issue and genetic crop improvement is seen as a solution to deal with this issue. While genetic editing techniques and genome mapping technologies are blooming, they depend on a similar improvement in phenotyping, since they are key to analyse plant responses to environmental characterization. In recent years, high-throughput and high-resolution phenotyping tools have made impressive progress and can now help relieving the current phenotyping bottleneck (Fiorani & Schurr 2013; Furbank & Tester 2011; Tardieu *et al.* 2017). Different phenotyping platforms are emerging around the world. They range from high-precision platforms for cell and organ characterization (Vargas *et al.* 2006) to multi-environment networks of fields, exploiting remote sensing (Virlet *et al.* 2017). At all scales, phenotyping facilities display spatial heterogeneity that needs to be separated from the genetic signal. For example, the spatial variability of incident light raises up to 30% between pots within a glasshouse or a growth chamber (Cabrera-Bosquet *et al.* 2016). There are also evidences of microclimate variations in greenhouses experiments (Brien *et al.* 2013). Therefore, correcting for spatial trends and using appropriate experimental designs is crucial for a precise estimation of genetic effects. Hence, the existing design and modelling theory for field experiments needs to be adapted for the phenotyping platforms.

[1] An extensive list of all the needed definitions is available in the glossary in the forepart of this thesis.

2.2 Experimental design in field trials

Experimental field trials in agriculture have always been affected by soil heterogeneity. As Van Es (2002) explains, soil is a continuum with variability on multiple scales. The heterogeneity is as much affected by microscopic interactions as by field-sized effects. Therefore, agricultural trials have always heavily relied on randomisation, blocking and replication to account for spatial variability and remove bias from the estimation of the treatment effects (Atkinson & Bailey 2001). For randomisation to be truly effective, stationarity of the mean and spatial independence assumptions need to be verified. Several studies have proven that it is rare that both these assumptions hold in field trials (Davidoff *et al.* 1986; Iqbal *et al.* 2005; Nielsen *et al.* 1973). Moreover, Van Es (1993) showed that even randomized designs can still be problematic for experiments with large numbers of treatments and low numbers of replications in the presence of spatial autocorrelation. A new class of design has been proposed involving the use of replicated plots for a percentage of the test lines: the "p-rep" designs (Cullis, Smith, *et al.* 2006; Velazco *et al.* 2017). Local field trends can influence groups of treatments in specific blocks. As a solution, several authors (Fagroud & Van Meirvenne 2002; Watson 2000) have suggested considering the spatial trends and autocorrelation structures when creating the design, by using prior soil information, but taking into consideration spatial variability in the design of a trial not only require previous information on the plot but is often costly and cumbersome. Furthermore, in practice, most experimenters have neither the capacity to implement advanced designs (in terms of computation power and statistical training), nor the capacity to analyse them. Finally, Van Es *et al.* (2007) showed that completely randomized (43 % in greenhouse trials) and random block designs (70 % in field trials) are still widely used. Considering this global issue, finding and using an appropriate design is complex task.

2.3 Spatial modelling for field trials

In order to increase the precision of the estimation of genetic effects, experimental designs need to be complemented with appropriate models of analysis. Mixed model analyses using the autoregressive ($AR1$) functions (Cullis & Gleeson 1991) have become a standard strategy in field trials. However, H. Piepho *et al.* (2015) recently discussed several issues with this model and have therefore proposed the use of the linear variance (LV) model (E. R. Williams & Luckett 1988) instead. More specifically, H. P. Piepho & E. R. Williams (2010) have proposed a revised version of this model, augmenting it into two dimensions ($AR1 \times AR1$) The main novelty resides in the addition of spatial components to a classic rows-columns model. Recently, M. X. Rodríguez-Álvarez, Boer, van Eeuwijk & P. H. Eilers (2018) introduced a novel spatial model that adjusts for both global and local trends simultaneously: the SpATS model (Spatial Analysis of field Trials with Splines). The new spatial method makes use of penalized splines (P. H. C. Eilers & Marx 1996) to estimate a bivariate smooth function over the rows and columns of a plot. Using the work of Lee Hwang (2010), Lee & Durbán (2011), and Lee, Durbán & P. Eilers (2013) the spatial variability is characterized using tensor products of two-dimensional P-splines (Dierckx 1995) and decomposed in a PS-ANOVA system. By exploiting the similarities between P-splines and mixed

models (Currie & Durban 2002; Durban *et al.* 2001; Wand 2003), the P-splines are expressed as a mixed model, which allows the use of classical mixed-model software but also the use of additional random and fixed effects to the model to better capture the variation along the 2-dimensional field. It has already been tested on simulated data (M. X. Rodríguez-Álvarez, Boer, van Eeuwijk & P. H. Eilers 2018) and previous field trials data (Lado *et al.* 2013) and showed promising results.

As Wilkinson *et al.* (1983) and Gilmour *et al.* (1997) highlight, in field trials data modelling, three main sources of spatial variations need to be accounted for:

Stationary variations [2] Large scale trends across the field (e.g. fertility trend, depth of soil, moisture)

Non-stationary variations Also natural variations but localized on part of the field (e.g. patch of soil moisture)

Extraneous variations Variations unrelated to a natural process, often due to the way the field is prepared (e.g. tillage, sowing practices, etc...)

A part of these variations can be attributed to systematic effects, e.g. sowing or planting, other to random effects such as fertility trends. While systematic effects can easily be modelled using factors and row-columns attributes, it is not case the case for random spatial variation. They are harder to model because there are no covariates to relate it to. Since the spatial variation has both random and systematic components, it makes sense to use the mixed model framework.
There are two main approaches to model spatial trends: one based on spatial variance-covariance structures; and the other based on smoothing techniques. In this thesis, the data extracted from the phenotyping platform are modelled using these 2 different models.

2.4 Thesis objectives

This master thesis falls within the scope of the second activity of the European project EPPN2020[3]. It is a research infrastructure project funded by Horizon 2020, that will provide access to 31 key plant phenotyping installations. It defines three research activities: (1) novel technologies and methods for environmental and plant measurements, (2) innovative design and analysis of phenotyping experiments across multiple platforms and (3) a European plant phenotyping information system. The project revolves around data acquisition, data analysis and data networking, so that every

[2] Risser (2016) defines a stationary process as follows:
Let C be a spatial covariance function, it is said to be stationary if the features of C do not depend on spatial location. More formally, a process $\{Y(\mathbf{s}) : \mathbf{s} \in G\}$ is said to be second-order stationary (or weakly stationary) if the following two properties hold:

1. $E[Y(\mathbf{s})] = E[Y(\mathbf{s}+\mathbf{h})] = c$ for some constant c and
2. $C(\mathbf{s}, \mathbf{s}+\mathbf{h}) = C(\mathbf{0}, \mathbf{h})$ for some spatial lag $\mathbf{h} \in \mathcal{R}^d$.

[3] European Plant Phenotyping Network 2020 `https://eppn2020.plant-phenotyping.eu/`

platform uses common, standardized practices and analysis protocols, that have been tested for robustness and quality.

The main goal was to assess the utility of statistical designs and mixed models to identify and correct for spatial trends (heterogeneity) in an aeroponic root installation at UCLouvain (Louvain-la-neuve). The idea is to set up an experiment in this installation using different genotypes (plant varieties) and a custom experimental design to account for possible complex environmental variations. It will be created using JMP, taking into account the specifities of the platform and the number of genotypes used. This approach allows the design to fit the experiment properly and avoids having to use a pre-made design, not optimal for the experiment. After data collection and image analysis, two different models will be used to model the spatial variability and to assess the quality of spatial prediction. The first one is a two-dimensional version of the linear variance model, revised by H. P. Piepho & E. R. Williams (2010). The second one is the SpATS (Spatial Analysis using Tensor product of Splines) model, recently created by M. X. Rodríguez-Álvarez, Boer, van Eeuwijk & P. H. Eilers (2018). The two models will be compared in term of their ability to estimate genotypic effects and to quantify spatial variability. These comparisons will be made using classical indicators (RMSE, ...) and other indicators, specific to spatial models for field trials (Oakey *et al.* 2006).

The experiment took place in February in the UCLouvain greenhouses. The installation consists of two aeroponic tanks of 495 plants located in a 64 m^2 greenhouse. Plants are held on strips, 5 plants per strip, 99 strips per tank. The specificity of the platform is that the plant rotate constantly so that their root system can be photographed every two hours. The experiment lasted 3 weeks, after which the plants became too large for the platform. The experiment included two tanks. In the first one, plants constantly moved to be pictured every 2 hours (usual set-up on this platform). In the second one, plants moved twice or three times a day to be pictured. This allowed comparing the effect of moving versus non-moving plants, which is a feature often available in the phenotyping platforms but poorly evaluated so far.

Since the UCLouvain platform focuses on the analysis of the root system, the main variable of interest in the experiment is the overall growth of the root system of each plant. An experiment generates a large amount of data in a raw, image-based format (approximately 200K images). Images allow temporal decoupling, provide a condensed set of information and are multi-dimensional (2D usually, but 3D scanning platforms are developping (Mooney *et al.* 2012)). A lot of tools are available for data analysis in phenotyping platforms (Lobet, Draye & Périlleux 2013). This makes the choice complicated for an external user, especially since most of these softwares are designed for a single specific purpose. Another challenge in root system architecture (RSA) characterisation is the inherent complexity of the system. Different techniques have been developed to best characterize the RSA in a cost-efficient way (Lobet & Draye 2013; Pound *et al.* 2013). Scientists of the UCLouvain platform have developed pipelines[4] that allow easy processing of the images captured in the platform to extract

[4]Here, pipelines are defined as computer programs designed to analyse raw data from phenotyping platforms.

quantitative root architecture information for the spatial models (Lobet & Draye 2013; Lobet, Pagès, *et al.* 2011).

This thesis can be summarised in four main points: create an appropriate experimental design for a phenotyping experiment, analyse data from a high-throughput platform, comparing the efficiency of various spatial models to correct for heterogeneous and non-linear spatial trends and developing the appropriate R scripts.

Chapter 3

Material and methods

3.1 Optimal experimental designs

Say how custom designs are more efficient because they fit to the problem rather than having to the change the problem to fit a design. The main principle in custom design is to maximize an optimality criterion to get the best design possible.

3.1.1 Orthogonal designs

In design of experiments, orthogonal designs are interesting because they guarantees that each main effect and interaction can be estimated independently. Meaning that the effect of one factor or interaction can be estimated separately from the effect of other factors and that the addition or subtraction of terms in the model does not affect the estimates. In a regression or ANOVA-type model, the best linear unbiased predictor (or BLUE) of the regression coefficients is obtained by using the ordinary least squares (OLS) method, because it minimizes the variance of the estimators. These variances are often represented in variance-covariance (VCOV) matrix, where the diagonal is the variance of the estimators and the non-diagonal elements are the pairwise covariances between estimators. The inverse of this matrix is the information matrix, because it summarizes the available information on the model's coefficients (a low variance means a lot of information, and inversely). As detailed in Goos & Jones (2011), when the information matrix is diagonal, then the design is said to be orthogonal.

3.1.2 Optimality criteria

In order to generate optimal designs, one needs to use an optimality criterion to compare different designs. The two main criteria are the D-optimality and the I-optimality. The first one aims at minimizing the variance of the factors effects estimates and is more useful for significance testing. D-optimal designs are therefore more appropriate for screening experiments. The second one aims at minimizing the average relative prediction variance over the experimental region. I-optimal designs are focused on prediction and thus are more suited to response surface experiments. There also exists a G-optimality criterion that is similar to the I-optimality criterion but minimizes

the maximum prediction variance. Recent work (Rodríguez *et al.* 2010) has shown that I-optimal designs are often better choices than the G-optimal ones. Since this is a screening experiment, only the D-optimality is detailed here. For more information about I-optimal and G-optimal design, refer to Atkinson (2014) and Goos & Jones (2011).

D-optimality

As said previously, for an orthogonal design all the non-diagonal elements of the VCOV matrix are null, and thus, the determinant is simply the product of all the diagonal elements, i.e. the estimators variances. Since the goal is to have the smallest variance of the estimates, the VCOV matrix with the smallest determinant will have the estimates with the smallest variance. Minimizing the determinant of the VCOV matrix is similar to maximizing the determinant of the information matrix. Therefore, the design with factor settings that maximize the determinant of information matrix, will maximize the available information about the model's parameters. This design is called the "D-optimal design", where the "D" stands for determinant and the value of the determinant itself is called the "D-optimality criterion".

For any model with two-levels factors and two-factor interaction effects, orthogonal designs will always be D-optimal. However if the number of runs is not a multiple of 4 then there are no orthogonal designs available for two-level factors. This condition offers little flexibility for experimenters and is not always feasible. In contrast, the optimal experimental design approach allows for any number of runs. However, in non-orthogonal designs the variance of the estimates is inflated and the estimates are correlated. Nevertheless this inflation is usually small and the correlation is too small to cause any concerns. Therefore there exist non-orthogonal designs that still maximize the information of the model being estimated. The D-optimal designs may not be unique. For a specified number of runs, several designs might have the maximal value for the determinant of the information matrix.

3.1.3 Generating optimal designs

In order to generate an optimal design, the determinant of the $\mathbf{X}'\mathbf{X}$ matrix needs to be computed multiple times. Therefore algorithms are used to gain time and avoid errors. Several algorithms exist but the most common one is the coordinate exchange algorithm, created by Meyer & Nachtsheim (1995). It has the advantage to run in polynomial time, which means that the time needed to find an optimal design does not explode when the size of the design increases. Another similar algorithm is the point-exchange algorithm, created by Fedorov (1972) and modified several times to speed it up (Atkinson & Donev 1989; Johnson & Nachtsheim 1983). The main drawback of this algorithm is that it needs a list of possible design points as input, which can be quite tedious to do for large designs. In recent years, other types of algorithm such as genetic algorithms (Heredia-Langner, Carlyle, *et al.* 2003; Heredia-Langner, Montgomery, *et al.* 2004), simulating annealing algorithms (Bohachevsky *et al.* 1986; Meyer & Nachtsheim 1988) and tabu search algorithms (Jung & Yum 1996) have been used in experimental designs. While these algorithms maintain a level of performance

comparable to more traditional design construction techniques, they are not as popular because they are either far more complex, only feasible in some specific cases or better for some specific models and do not lead to designs that make a significant difference in practice.

Coordinate-exchange algorithm

The coordinate-exchange algorithm proceeds by iterating through the rows of the matrix of factors settings

$$\mathbf{D} = \begin{bmatrix} x_{11} & x_{21} & \dots & x_{k1} \\ x_{12} & x_{22} & \dots & x_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1n} & x_{2n} & \dots & x_{kn} \end{bmatrix}, \tag{3.1.31}$$

called the design matrix, of an experiment with n runs and k factors. The lines of this matrix essentially represent the coordinates of the runs in the experimental space, where each factor of the experiment is a dimension. This algorithm is called the coordinate-exchange algorithm, because in each iteration of the algorithm, possible changes for every element of the design matrix are considered.
It is straightforward to see that the design matrix $\mathbf{D}$ is a submatrix of the model matrix $\mathbf{X}$. It can happen the D-optimality criterion is zero. In those cases, the design is called singular and the inverse of the $\mathbf{X}'\mathbf{X}$ matrix does not exist. To avoid singularity, the number of design points (different rows in the design matrix $\mathbf{D}$) must be greater than or equal to the number of model parameters.

The algorithm starts by generating a random design. For all continuous factors, the algorithm generates random values on the interval $[-1, +1]$. For all factors that are categorical, the algorithm randomly chooses a value in a discrete set of levels. This random starting design is almost always non-singular. If the design happens to be singular, then another new random starting design is computed.
In the next step, the algorithm improves the design on an element-by-element basis. For each element of the starting design, x_{ij}, a change to either -1 or $+1$ is considered, and its impact on the D-optimality criterion is evaluated. The change that increases the value of the D-optimality criterion the most, is kept. After investigating changes in each element of the design, the process is repeated until no element changes within an entire iteration through the factor settings or until a prespecified maximum number of iterations is reached. The obtained design is the best among a set of neighbouring designs but it is often a locally optimal design that is different for each random starting design. To select the best among all locally optimal designs, the algorithm is repeated a large number of times. The globally optimal design is then selected among all the locally optimal ones, as the one that yields the highest D-optimality criterion.

3.1.4 Generating the design

A custom design was created for the experimental set-up of the phenotyping platform, where the goal was to quantify the genotype and tank effect. Four factors were considered:

Tank In which tank was the plant situated (moving or still).

Strip Which of the 99 strip was used (1 to 99).

Position What was the position on the strip (1 to 5).

Genotype Which one of the 30 genotypes was used (1 to 30).

To fit the design, the design of experiment (DOE) tool was used in JMP. The four categorical factors were specified and *Tank* and *Strip* were set to "very hard to change" and "hard to change", respectively. Two whole plots of 99 sub-plots each were specified to match the tanks and the strips. With 99 strips of five positions inside two tanks, 990 experimental runs were available.
Initially the design was supposed to take into account the 99 different strips individually but the program couldn't converge to an optimal design because of its complexity. Instead only 33 strips were considered and the design was replicated 3 times to match the number of runs. Figures 3.1 and 3.2, display a schematic view of the design for the moving and still tank respectively.
The seeds were provided by the national institute of agronomic research (INRA) in Montpellier, France, as part of their own research on these historical series[1]. They sent a total of 30 seeds per genotype and besides that, an extra 150 seeds of another genotype, called the "border genotype", was sent by the provider. The main utility of the border genotype is to fill the gaps left by non germinated seeds. Since, in this experiment, only 900 runs (30 genotypes x 30 seeds) were occupied, the border genotype was also be used to fill the 90 empty runs. Since this genotype is not part of the historical series, it was not considered in the design of the experiment.

3.1.5 Design's characteristics

[1]Historical series correspond to varieties that have been cultivated and bred for some time, mainly due to their physiological specificities.

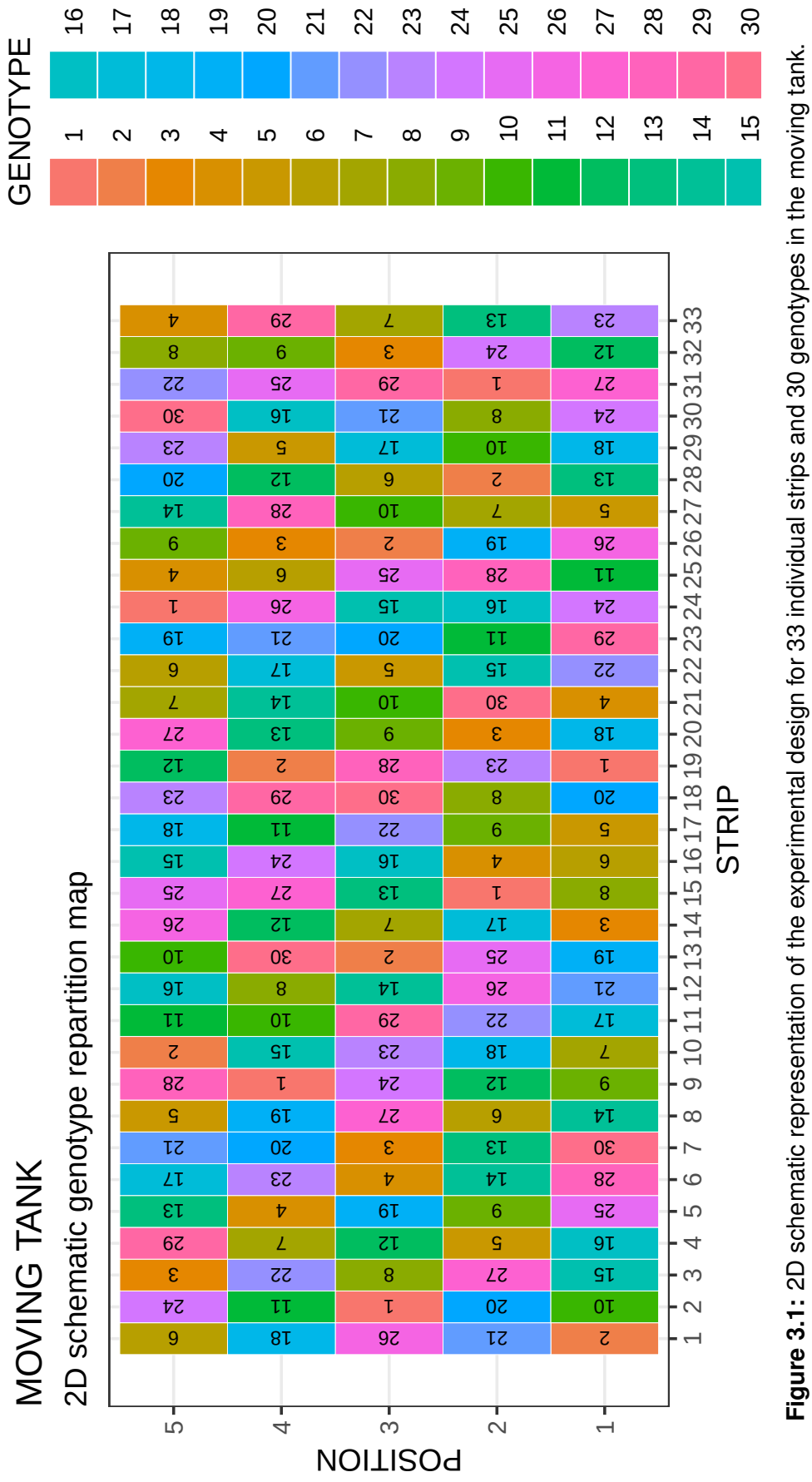


Figure 3.1: 2D schematic representation of the experimental design for 33 individual strips and 30 genotypes in the moving tank.

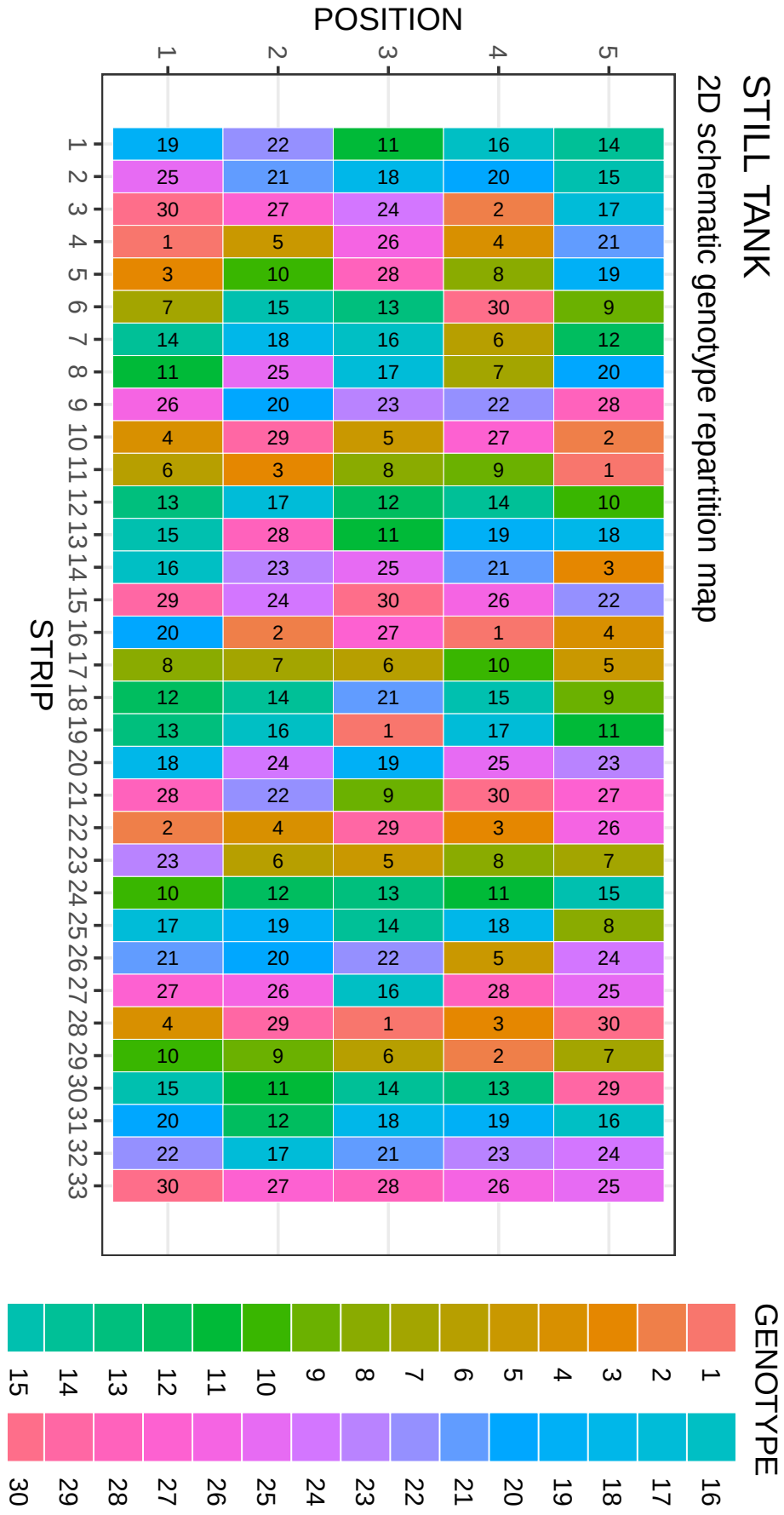


Figure 3.2: 2D schematic representation of the experimental design for 33 individual strips and 30 genotypes in the still tank.

3.2 Phenotyping experiment

The phenotyping experiment took place between February 25th and March 13th. The seeds were first germinated and then transferred onto the platform. After the end of the experiment the plants were weighted, dried and weighted again to obtain dry and fresh weight.

3.2.1 Germination

Previous experiments in the greenhouses showed that germination of maize seeds on the platform often lead to asphyxiation of the seeds. Because of this, the seeds were germinated in an outside germination chamber and were only transferred onto the platform once germinated.

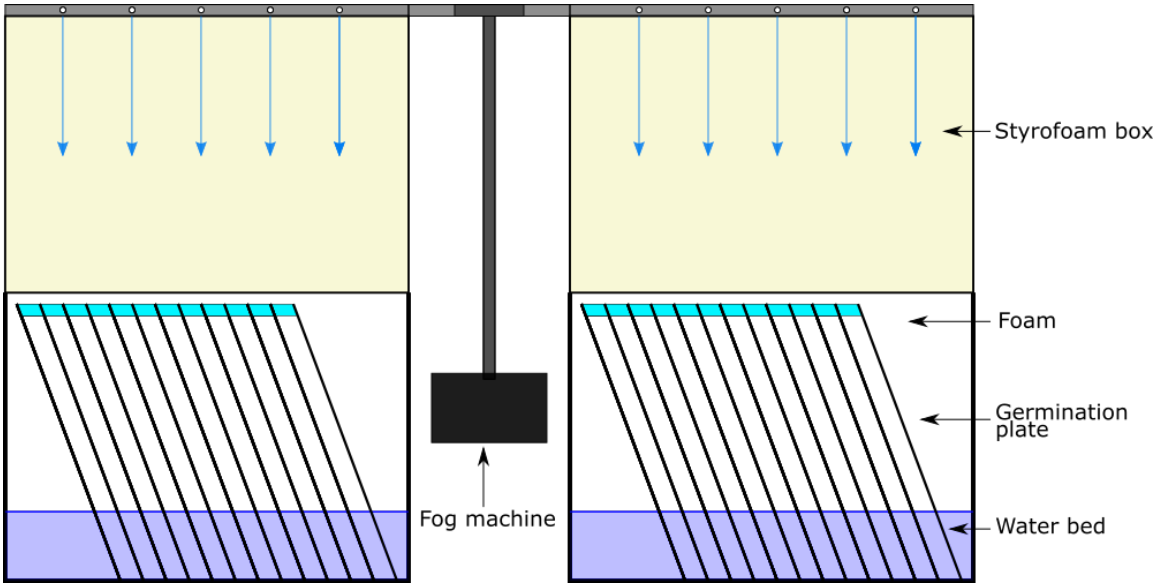
The seeds were placed in a temperature-controlled room at 20°C for 3 days, inside a germination chamber. The chamber consisted of 2 PVC trays to which an air-fog machine was connected, to keep the seeds moist. Inside each tray, PVC plates were disposed diagonally and evenly spaced (fig. 3.3a). On those plates, the seeds were arranged on a filtering paper sheet with ledges to support the weight of the seeds (figure 3.3b and figure 3.3c). The bottom of the trays were filled with water to keep the filtering paper moist. There were 17 plates in total, 15 for the 30 genotypes and 2 for the border genotype (150 seeds dispatched on 2 plates).

After 3 days into the chamber, not all seeds were germinated, table 3.1 presents the germination rates and mean seed weights for all the genotypes used (including the border genotype). The non-germination was mainly due to the fact that seeds fell into the water bed and because mold grew on some filtering paper.

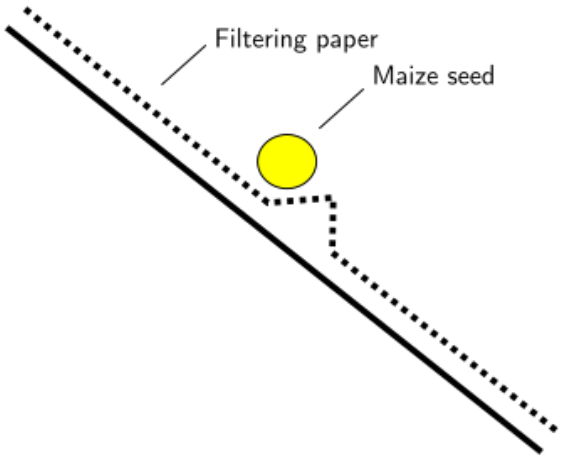
3.2.2 Phenotyping platform

The phenotyping platform is located inside a greenhouse in the facilities of the UCLouvain (Louvain-la-Neuve, Belgium). It consists of two aeroponic tanks on which are arranged 99 styrofoam strips, each with five holes. At the end of each tank is a high definition camera that scans the root system of each plant individually, when it passes in front of it (fig. 3.4a). The strips rotate in a clockwise fashion in the tank and a full rotation is completed in 2 hours. Three sprinklers are placed regularly at the bottom of each side of the tank (fig. 3.4b). The sprinklers spray nutrient solution[2] at regular intervals, set by the operator. The spraying pattern (interval and duration) can be differentiated between day and night and can be modified at any moment of the experiment. In this case the patterns were 5 seconds of spraying ever 295 seconds all the time. During the experiment, the temperature of the greenhouse was set to 20°C at day and 18°C at night and the lights were on from 6 AM to 10 PM. At the start of the experiment seeds were placed inside a foam cork and then placed inside a hole on a strip (fig 3.4c). They were placed at the bottom to allow the root system to grow freely. The corks are drilled vertically to let the leaf system develop with less resistance and allow a direct access to sunlight. More information about the platform is available in appendix C.

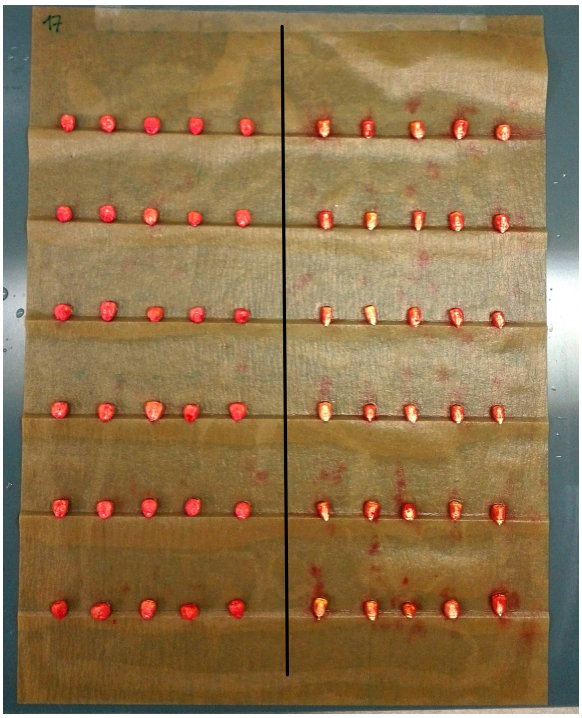
[2]The precise concentration of the Hoagland solution is presented in appendix B



(a) Global schematic view of the germination chamber: a fog machine assure constant humidity in the germination chambers by creating fog at regular intervals (the blue arrows represent the path of the fog). The plates are placed at a 60° angle and 5 cm apart



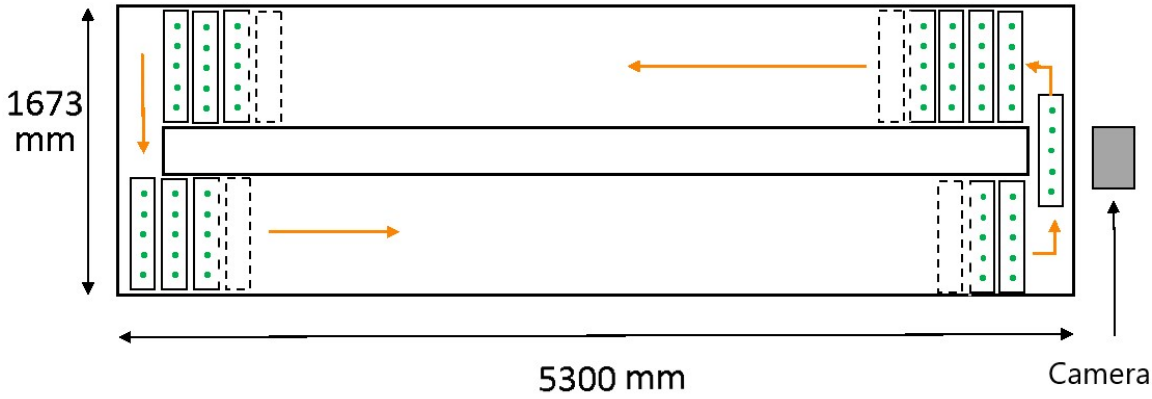
(b) Schematic view of a germination ledge on a PVC plate: each seed is fixed in position on the ledge by an additional drop of agar solution to avoid any fall-off



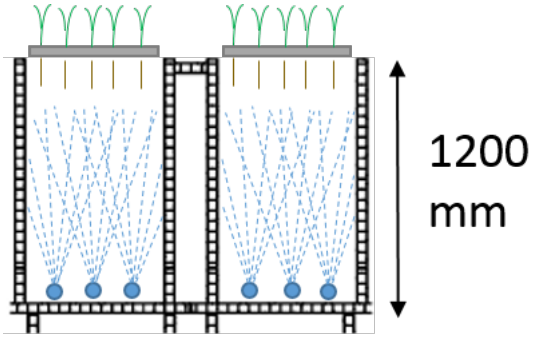
(c) 30 cm by 40 cm PVC plate with seeds on filtering paper (the black line represents the separation between the two genotypes on the plate). Each sheet had 6 rows of 10 seeds with one genotype on the left and one genotype on the right.

Figure 3.3: Germination chamber diagram with detailed view and pictures

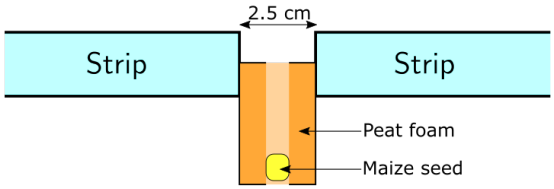
After 3 days in the germination chamber, the germinated seeds were transferred onto the platform following the created design, and the non-germinated seeds were discarded. The first tank was constantly moving but only scanning the root systems once a day, while the second tank was only moving once a day to scan the root system of the plants and stayed still the rest of the day.



(a) Schematic view of an aeroponic tank: plants are hold on strips, 5 plants per strip (green dots on layout). There are 99 strips in the tank for a total of 495 plants/tank. Strips move in the direction indicated by orange arrows



(b) Transversal schematic view of an aeroponic tank of the platform: at the bottom of each tank, sprinklers (represented in blue in the layout) are disposed at regular interval and spray nutritive solution



(c) Close up schematic view of a strip: inside each hole, seeds are placed inside a pierced peatfom cork to allow the root system to develop frelly

Figure 3.4: Detailed diagrams about the phenotyping platform

Table 3.1: Germination rate and mean seed weight for each genotype used. (there is no data concerning the germination rate of the border genotype because it was not measured).

Genotype	Germination rate (%)	Mean seed weight (g)
1	80.00	0.28
2	86.67	0.36
3	96.67	0.36
4	73.33	0.28
5	100.00	0.32
6	96.67	0.24
7	96.67	0.19
8	70.00	0.31
9	96.67	0.33
10	96.67	0.25
11	60.00	0.33
12	93.33	0.27
13	90.00	0.24
14	86.67	0.31
15	56.67	0.35
16	90.00	0.28
17	90.00	0.30
18	86.67	0.26
19	100.00	0.26
20	86.67	0.28
21	86.67	0.32
22	53.33	0.30
23	73.33	0.28
24	100.00	0.16
25	96.67	0.19
26	96.67	0.25
27	96.67	0.28
28	83.33	0.36
29	93.33	0.30
30	73.33	0.35
31	/	0.38

3.3 Data processing

After 15 days, the plants were considered fully grown and the experiment was stopped. The leaf and root systems were separated and weighted individually on scales precise to 0.001 g. They were then dried for 3 days in a 70°C oven. After the drying process, they were weighted again. For each plant, the remaining of the seed was consistently kept on the root system. Two kind of data were obtained from the experiment: weight data (dry and fresh) of the fully grown plants and root scan data. Five variables were kept for the spatial analysis:

- $FRESH\_RS$: fresh weight of the root system
- $FRESH\_LS$: fresh weight of the leaf system
- DRY_RS: dry weight of the root system
- DRY_LS: dry weight of the leaf system
- $AREA$: percentage of the total area occupied by the root system

To be used as inputs in the models, those data need to be processed. The following sections details this step.

3.3.1 Weight data

For some plants, the germinated seeds placed on the platform did not fully grow or had an abnormal growth, but all the plants were still weighted to avoid leaving out any data. Therefore some data points need to be handled more carefully, as they do not represent the genotype's growth correctly. However, a correct growth, representative of the genotype, is hard to define because the influence of the conditions on each genotype is unknown. Therefore, instead of choosing which plants are outliers in a binary way, we attributed weights to each plant to express the quality of the data. Those weights were established by reviewing the final root scan of each plant and checking the different factors that could have an influence. The factors chosen are the following:

- NO_RS: no additional root to the primary root
- NO_LS: no visible leaf system
- BAD_LS: leaf system grew under the strip (or abnormally in general)
- NO_SEED: no seed (or an empty cork) present on this position
- NOT_FG: plant not fully grown
- $OVERLAP$: leaf (or root) system of another plant overlaps on the root scan
- OK: no influential factors

An visualization of those factors with example pictures is presented in figure 3.5. Some plants were attributed several influential factors, but OK, NO_SEED and NOT_FG were considered as exclusive. The factor attribution was ambiguous for some pictures, in those cases the plant were considered OK to avoid losing any data points. Following the determination of the factors, weights were attributed to each variable according to the weight matrix, presented in table 3.2.

Table 3.2: Weight attribution matrix for the different factors and variables. LS weight is both the fresh and dry weight for leaf system and RS weight is the same for the root system

CODE	LS weight	RS weight	Area
NO_RS	1	1	1
NO_LS	2	2	2
BAD_LS	3	3	3
NOT_FG	4	4	4
NO_SEED	0	0	0
OK	5	5	5
OVERLAP	5	5	0

3.3.2 Root pictures

The area of the root system was computed using the final root scan of each plant. Each picture was converted to gray-scale, resized to a specific bounding rectangle and binarized. The bounding box is illustrated on figure 3.5g. The threshold for the binarization was 140 on a 0 to 255 scale. Since the roots are black on the original picture, the binarization was completed without any issues. The area of the root system was expressed as a percentage of the total area of the bounding box, by counting the black pixels in the image. All the processing was done in python using the opencv package.

(a) NO_RS: no additional root, other parent leaf system than the primary

(b) NO_LS: no apparent leaf system

(c) BAD_LS: the leaf system grew under the strip and deviated to the right (obstruction of another plant)

(d) NO_SEED

(e) NOT_FG: both system are present but underdeveloped compared to images

(f) $OVERLAP$: the leaf system of another plant hides the root system of the plant

(g) OK, the green box represents the bounding area for the the computation of the root area

Figure 3.5: Example pictures of the influential factors chosen for the weight attribution.

3.4 SpATS model

In this section, the SpATS model is introduced. For a more thorough treatment of the model and all its components, see M. X. Rodríguez-Álvarez, Boer, van Eeuwijk & P. H. C. Eilers (2016).
Consider a field trial of n plots arranged in a rectangular grid, where the plot positions are collected in vectors of row $(\mathbf{r})$ and column $(\mathbf{c})$ coordinates. If $\mathbf{y}$ is the vector of plot data in field order, a common model for $\mathbf{y}$, to use as a starting point is

$$\boldsymbol{y} = \mathbf{1}_n\beta_0 + \boldsymbol{Z}_r\boldsymbol{c}_r + \boldsymbol{Z}_c\boldsymbol{c}_c + \varepsilon \tag{3.4.01}$$

were $\mathbf{1}_n$ is a column-vector of ones of length n, $\boldsymbol{c}_r$ and $\boldsymbol{c}_c$ are, respectively, the random effect coefficients for the rows and columns and associated matrix $\boldsymbol{Z}_r$ and $\boldsymbol{Z}_c$. To fully capture complex spatial patterns, a smooth bivariate surface jointly defined over the row and column positions is added to the model, which becomes

$$\boldsymbol{y} = f(\boldsymbol{u}, \boldsymbol{v}) + \boldsymbol{Z}_r\boldsymbol{c}_r + \boldsymbol{Z}_c\boldsymbol{c}_c + \boldsymbol{\varepsilon} \tag{3.4.02}$$

where $\mathbf{u}$ and $\mathbf{v}$ are, respectively, the vector of row and columns positions and where $f(.,.)$ represents the smooth bivariate function. Note that the intercept term, β_0 is embedded into $f(u, v)$. To better understand this function, let us decompose it in a nested-ANOVA structure

$$\begin{aligned} f(\boldsymbol{u}, \boldsymbol{v}) = & \underbrace{\mathbf{1}_n\beta_0 + \boldsymbol{u}\beta_1 + \boldsymbol{v}\beta_2 + \boldsymbol{u} \odot \boldsymbol{v}\beta_3}_{\text{Bilinear polynomial}} \\ & + \underbrace{f_u(\boldsymbol{u}) + f_v(\boldsymbol{v}) + \boldsymbol{u} \odot h_v(\boldsymbol{v}) + \boldsymbol{v} \odot h_u(\boldsymbol{u}) + f_{u,v}(\boldsymbol{u}, \boldsymbol{v})}_{\text{Smooth part}} \end{aligned} \tag{3.4.03}$$

where $\odot$ denotes the element-wise matrix product[3]. There are now two components to the model: a bilinear polynomial part(parametric) and a smooth part (non-parametric). The parametric part includes the linear trends along rows (β_1) and columns (β_2) as well as a linear interaction trend (β_3). The smooth part models the deviation from the compound linear trend, and can be decomposed in the following elements:

- $f_u(\boldsymbol{u})$ is a smooth trend along the rows, identical for all columns (i.e., a main smooth effect).
- $f_v(\boldsymbol{v})$ is a smooth trend along the columns, identical for all rows.
- $\boldsymbol{v} \odot h_u(\boldsymbol{u})$ and $\boldsymbol{u} \odot h_v(\boldsymbol{v})$ are linear-by- smooth interaction trends. For instance, $\boldsymbol{u} \odot h_v(\boldsymbol{v})$ is a varying coefficient surface trend, consisting of functions, linear in the rows, for each column, but with slopes that change smoothly along the columns, $h_v boldsymbol v)$.
- $f_{u,v}(\boldsymbol{u}, \boldsymbol{v})$ is a smooth-by-smooth interaction trend jointly defined over the row and column directions.

[3] See appendix A for details about the element-wise matrix product.

In total, six components are used to model the surface f. This may seem like a lot but this allows the translation of model 3.4.02 into a standard mixed model. An interesting property of this proposal is that since u and v are row and column position, it allows depicting the spatial trend in a grid finer than the number of rows and columns. Figure 3.6 shows an example of those six components in the context of a barley uniformity performed by K. A. Williams (1988). It shows clearly how the additional components, help to capture small variations in the spatial data.

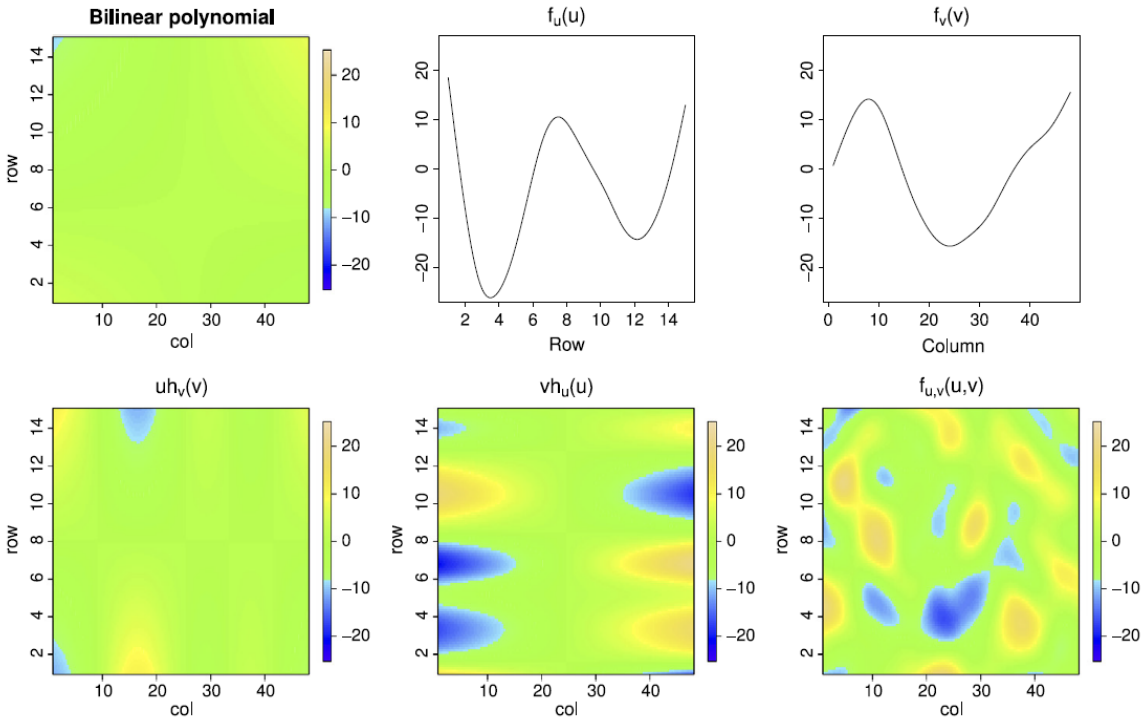


Figure 3.6: Bilinear and smooth components of the PS-ANOVA decomposition of the estimated spatial trend for the barley uniformity data from M. X. Rodríguez-Álvarez, Boer, van Eeuwijk & P. H. Eilers (2018).

3.4.1 Modelling using P-splines

The functions f_u, f_v, h_u and h_v are constructed with variations on one-dimensional P-splines, while $f_{u,v}$ is based on tensor products P-splines.
For clarity's sake, let us consider a model only containing a smooth bivariate surface and an error term

$$y_i = f(u_i, v_i) + \varepsilon_i, \text{ with } \varepsilon_i \sim N\left(0, \sigma^2\right). \tag{3.4.11}$$

Lee, Durbán & P. Eilers (2013) show that it can be represented using B-splines[4]. Let us form two B-splines basis:

1. one for the columns, $\hat{\boldsymbol{B}}$ with $b_{il} = \hat{B}_l(u_i)$, where $\hat{B}_l(u_i)$ is the lth B-spline of the basis, evaluated at u_i

[4]See appendix A for details about B-splines and P-splines.

2. and one for the rows, $\breve{\boldsymbol{B}}$ with $b_{ip} = \breve{B}_p(v_i)$, where $\breve{B}_l(v_i)$ is the pth B-spline of the basis, evaluated at v_i.

Then, the smooth-by-smooth interaction can be written using those basis

$$f(u_i, v_i) = \sum_{l=1}^{L} \sum_{p=1}^{P} \hat{B}_l(u_i) \breve{B}_p(v_i) \alpha_{lp}, \tag{3.4.12}$$

where $\boldsymbol{\alpha} = (\alpha_{11}, \ldots, \alpha_{1P}, \ldots, \alpha_{LP})^t$ is a vector of unknown regression coefficients of dimension $(LP \times 1)$. Note that $\hat{\boldsymbol{B}}$ and $\breve{\boldsymbol{B}}$ are matrices of order $n \times L$ and $n \times P$ respectively, where L and P are the number of the B-spline basis functions. Dierckx (1995) shows that, in the P-spline framework, the smooth-by-smooth interaction $f(u_i, v_i)$ is modelled by the tensor product of B-splines bases. Then, we can write, in matrix notation,

$$\boldsymbol{B} = \hat{\boldsymbol{B}} \boxdot \breve{\boldsymbol{B}} = \left(\hat{\boldsymbol{B}} \otimes \mathbf{1}_L^t\right) \odot \left(\mathbf{1}_P^t \otimes \breve{\boldsymbol{B}}\right), \tag{3.4.13}$$

where the operation $\boxdot$ is defined in terms of the Kronecker product of two matrices (denoted by $\otimes$) and the element-by-element multiplication of two matrices (denoted by $\odot$)[5]. Therefore model (3.4.11) can be written in matrix notation:

$$\boldsymbol{y} = \boldsymbol{B}\boldsymbol{\alpha} + \boldsymbol{\epsilon}. \tag{3.4.14}$$

The coefficients of this parametric model can be estimated by minimizing the sum of squares. The explicit solution is then:

$$\hat{\boldsymbol{\alpha}} = (\mathbf{B}^t\mathbf{B})^{-1}\mathbf{B}^t\mathbf{y} \tag{3.4.15}$$

To prevent over-fitting, P. H. C. Eilers & Marx (1996) propose to incorporate a discrete penalty on the coefficient associated to adjacent B-splines. As described in details in appendix A.3, this penalty also determines the smoothness of the splines. The solution of equation 3.4.14 then becomes

$$\widehat{\boldsymbol{\alpha}} = \left(\boldsymbol{B}^t\boldsymbol{B} + \boldsymbol{P}\right)^{-1} \boldsymbol{B}^t\boldsymbol{y}, \tag{3.4.16}$$

where $\boldsymbol{P}$ is the penalty matrix. The details of this solution are presented in appendix A.4, but the important point to remember here, is that the smoothness of the bivariate surface is defined by the penalty matrix, which only depend on two tuning parameters $\hat{\lambda}$ (smoothing along the columns) and $\breve{\lambda}$ (smoothing along the rows).

3.4.2 Mixed model based smoothing parameter selection

As explained in M. X. Rodríguez-Álvarez, Boer, van Eeuwijk & P. H. C. Eilers (2016), $\mathbf{P}$ is rank-deficient, meaning that the rank is smaller than the number of rows and/or columns, and this causes numerical instability when applying mixed model estimation techniques. To obtain a full-rank penalty matrix, the key is to write model 3.4.14 as

$$\mathbf{B}\boldsymbol{\alpha} = \boldsymbol{X}_s\boldsymbol{\beta}_s + \boldsymbol{Z}_s\boldsymbol{c}_s. \tag{3.4.21}$$

[5]See appendix A for details about the Kronecker, and the element-wise products

There are now two bases: $\mathbf{X}_s$, with coefficients that are not penalized at all, and $\mathbf{Z}_s$, with a size penalty on its coefficients. This decomposition follows the proposal by Lee & Durbán (2011), based on eigenvalue decomposition which gives rise to a diagonal penalty matrix.
The two bases have the following structures:

$$\boldsymbol{X}_s = [\mathbf{1}_n, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{u} \odot \boldsymbol{v}] \quad \text{and} \quad \mathbf{Z}_s = [\mathbf{Z}_v, \mathbf{Z}_u, \mathbf{Z}_v \square \mathbf{u}, \mathbf{v} \square \mathbf{Z}_u, \mathbf{Z}_v \square \mathbf{Z}_u], \tag{3.4.22}$$

where $\boldsymbol{u}$ and $\boldsymbol{v}$ are still, respectively, the vectors of row and column positions. Here $\mathbf{Z}_u$ and $\mathbf{Z}_v$ are penalized version of the B-splines basis $\breve{\mathbf{B}}$ (rows) and $\hat{\mathbf{B}}$ (columns). This new way of writing the problem leads to another penalty matrix $\widetilde{\boldsymbol{P}}$, which is a block diagonal matrix. Each block of $\widetilde{\boldsymbol{P}}$ corresponds to a block in $\mathbf{Z}_s$. Similarly to $\boldsymbol{P}$, the penalty matrix of the previous section, this new penalty matrix only depends on the two tuning parameters $\hat{\lambda}$ (smoothing along the columns) and $\breve{\lambda}$ (smoothing along the rows). Figure 3.7 presents a diagram clarifying the structures and relations of the different matrices presented throughout this section.

This reformulation provides the ANOVA type decomposition discussed in the previous section (3.4.03), and explains how the bilinear smooth surface can be modelled using P-splines and tensor products of P-splines. The block structure of $\mathbf{X}_s$ and $\mathbf{Z}_s$ implies

$$\begin{aligned} f(\boldsymbol{u}, \boldsymbol{v}) &= \boldsymbol{X}_s \boldsymbol{\beta}_s + Z_s \boldsymbol{c}_s \\ &= \mathbf{1}_n \beta_0 + \boldsymbol{u} \beta_1 + \boldsymbol{v} \beta_2 + \boldsymbol{u} \odot \boldsymbol{v} \beta_3 \\ &+ \underbrace{f_v(\boldsymbol{v})}_{\boldsymbol{Z}_v \boldsymbol{c}_{s1}} + \underbrace{f_u(\boldsymbol{u})}_{\boldsymbol{Z}_u \boldsymbol{c}_{s2}} + \underbrace{\boldsymbol{u} \odot h_v(\boldsymbol{v})}_{[\boldsymbol{Z}_v \square \boldsymbol{u}] \boldsymbol{c}_{s3}} + \underbrace{\boldsymbol{v} \odot h_u(\boldsymbol{u})}_{[\boldsymbol{v} \square \boldsymbol{Z}_u] \boldsymbol{c}_{s4}} + \underbrace{f_{u,v}(\boldsymbol{u}, \boldsymbol{v})}_{[\boldsymbol{Z}_v \square \boldsymbol{Z}_u] \boldsymbol{c}_{s5}}, \end{aligned} \tag{3.4.23}$$

where $\boldsymbol{c}_{sk}$ $(k = 1, \ldots, 5)$ contains the elements of $\boldsymbol{c}_s$ that correspond to the kth block of $\boldsymbol{Z}_s$, i.e. $\boldsymbol{c}_s = (\boldsymbol{c}_{s1}^t, \ldots, \boldsymbol{c}_{s5}^t)^t$. The details about the specific block component of $\boldsymbol{Z}_s$ and the computation of the new penalty matrix are available in the paper of M. X. Rodríguez-Álvarez, Boer, van Eeuwijk & P. H. Eilers (2018) and the appendices therein.

Therefore, using this new notation, model 3.4.11 that only contains a smooth bivariate surface and an error term can be rewritten in the following way:

$$\boldsymbol{y} = \boldsymbol{X}_s \boldsymbol{\beta}_s + \boldsymbol{Z}_s \boldsymbol{c}_s + \boldsymbol{\varepsilon}, \text{ with } \boldsymbol{\varepsilon} \sim N\left(\mathbf{0}, \sigma^2 \boldsymbol{I}_n\right) \text{ and } \boldsymbol{c}_s \sim N\left(\mathbf{0}, \boldsymbol{G}_s\right), \tag{3.4.24}$$

where $\boldsymbol{G}_s = \sigma^2 \widetilde{\boldsymbol{P}}^{-1}$. It is straightforward to see that $\boldsymbol{G}_s$ also has a block diagonal structure, similar to that of $\widetilde{\boldsymbol{P}}$ (this structure is also represented on figure 3.7). However, $\boldsymbol{G}_s$ depends on two different parameters, $\breve{\sigma}^2 = \sigma/\breve{\lambda}$ and $\hat{\sigma}^2 = \sigma/\hat{\lambda}$, which are variances parameters. As shown in the diagram in figure 3.7, the same variance parameters control the smoothness of the both the main effects and interactions terms. This prevents the use of standard mixed models software for estimation since $\mathbf{G}_s$ has its last block depending on both $\breve{\sigma}^2$ and $\hat{\sigma}^2$, but in a non-linear way. Even though M. X. Rodríguez-Álvarez, Lee, *et al.* (2015) presented a specialized algorithm to deal with this issue, here the PS-ANOVA decomposition approach (Lee, Durbán & P. Eilers 2013) is used to allow the use of standard mixed model estimation procedures.Lee, Durbán & P. Eilers (2013) therefore propose to use a different variance component for each smooth

component in $\mathbf{G}_s$, thus redefining this matrix as a linear function of variance parameters:

$$\boldsymbol{G}_s = \bigoplus_{k=1}^{5} \boldsymbol{G}_{sk} = \text{ blockdiag } (\boldsymbol{G}_{s1}, \boldsymbol{G}_{s2}, \boldsymbol{G}_{s3}, \boldsymbol{G}_{s4}, \boldsymbol{G}_{s5}), \tag{3.4.25}$$

where $\boldsymbol{G}_{sk}$ is the kth block of the $\mathbf{G}_s$ matrix, depending on the specific variance component σ^2_{sk}. In other words, here the tensor product P-splines mixed model is represented as the sum of 5 sets of mutually independent Gaussian random components $\mathbf{c}_{sk}$, each depending on one variance σ^2_{sk} $(k = 1, \ldots, 5)$.
Within this mixed model framework, the smoothing parameters, defined earlier as the ratio between the residual variance and the corresponding variance effect $\lambda_{sk} = \sigma^2_e/\sigma^2_{sk}$, are determined by restricted maximum likelihood (REML). Therefore the smoothness of the spatial surface is tuned by five distinct parameters, applying anisotropic (direction-dependant) smoothing. This parametrization provides flexibility to account for both global and local variations in the field. Furthermore, the decomposition of $f(\mathbf{u}, \mathbf{v})$ enables a more explicit interpretation of the main patterns of spatial variation (M. X. Rodríguez-Álvarez, Boer, van Eeuwijk & P. H. Eilers 2018).

3.4.3 Spatial models for field trials

The tensor product P-spline, presented in the previous section, constitutes the base for the analysis of agricultural field trials because it allows the modeling of the random spatial variation typically presented in a field. On top of this spatial field, we need to build up a more complex models in order to account for the genetic variation, the different tanks, strips and positions. From now on, we therefore consider the following linear mixed model

$$\boldsymbol{y} = \underbrace{\boldsymbol{X}_s\boldsymbol{\beta}_s + \boldsymbol{Z}_s\boldsymbol{c}_s}_{f(\boldsymbol{u},\boldsymbol{v})} + \boldsymbol{X}_d\boldsymbol{\beta}_d + \boldsymbol{Z}_d\boldsymbol{c}_d + \varepsilon, \text{ with } \boldsymbol{c}_s \sim N(\boldsymbol{0}, \boldsymbol{G}_s) \text{ and } \boldsymbol{c}_d \sim N(\boldsymbol{0}, \boldsymbol{G}_d), \tag{3.4.31}$$

where $\boldsymbol{X}_s$, $\boldsymbol{Z}_s$ and $\boldsymbol{G}_s$ are defined in the previous section and form the mixed model expression of the smooth spatial surface, and $\boldsymbol{X}_d$ and $\boldsymbol{Z}_d$ represent column-partitioned matrices, associated respectively with fixed and random components. Since we do not have any check genotypic varieties or resolvable block effect, the only extra fixed effects are: an intercept (1_n) and the tank effect (t). We assume that the $\boldsymbol{X}_d$ matrix has full-rank. The position on the strip (p) and strip (s) variables are added as random effects in $\mathbf{Z}_d$. Therefore $\mathbf{X}_d = [\mathbf{X}_{1_n}, \mathbf{X}_t]$, $\boldsymbol{Z}_d = [\boldsymbol{Z}_{dp}, \boldsymbol{Z}_{ds}]$ and $\boldsymbol{G}_d =$ blockdiag $(\boldsymbol{G}_{dp}, \boldsymbol{G}_{ds})$. We further assume that $\boldsymbol{c}_s$ and $\boldsymbol{c}_d$ are independent. To keep the notation simple, we rewrite model (3.4.31) as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{c} + \varepsilon, \text{ with } \boldsymbol{c} \sim N(\boldsymbol{0}, \boldsymbol{G}) \text{ and } \varepsilon \sim N\left(\boldsymbol{0}, \sigma^2\boldsymbol{I}_n\right) \tag{3.4.32}$$

where $\boldsymbol{X} = [\boldsymbol{X}_s, \boldsymbol{X}_{d_1}, \mathbf{X}_{d_t}]$, $\boldsymbol{Z} = \left[\boldsymbol{Z}_s, \boldsymbol{Z}_{d_p}, \mathbf{Z}_{d_s}\right]$, and

$$\boldsymbol{G} = \text{ blockdiag } (\boldsymbol{G}_s, \boldsymbol{G}_{dp}, \boldsymbol{G}_{ds}) \tag{3.4.33}$$

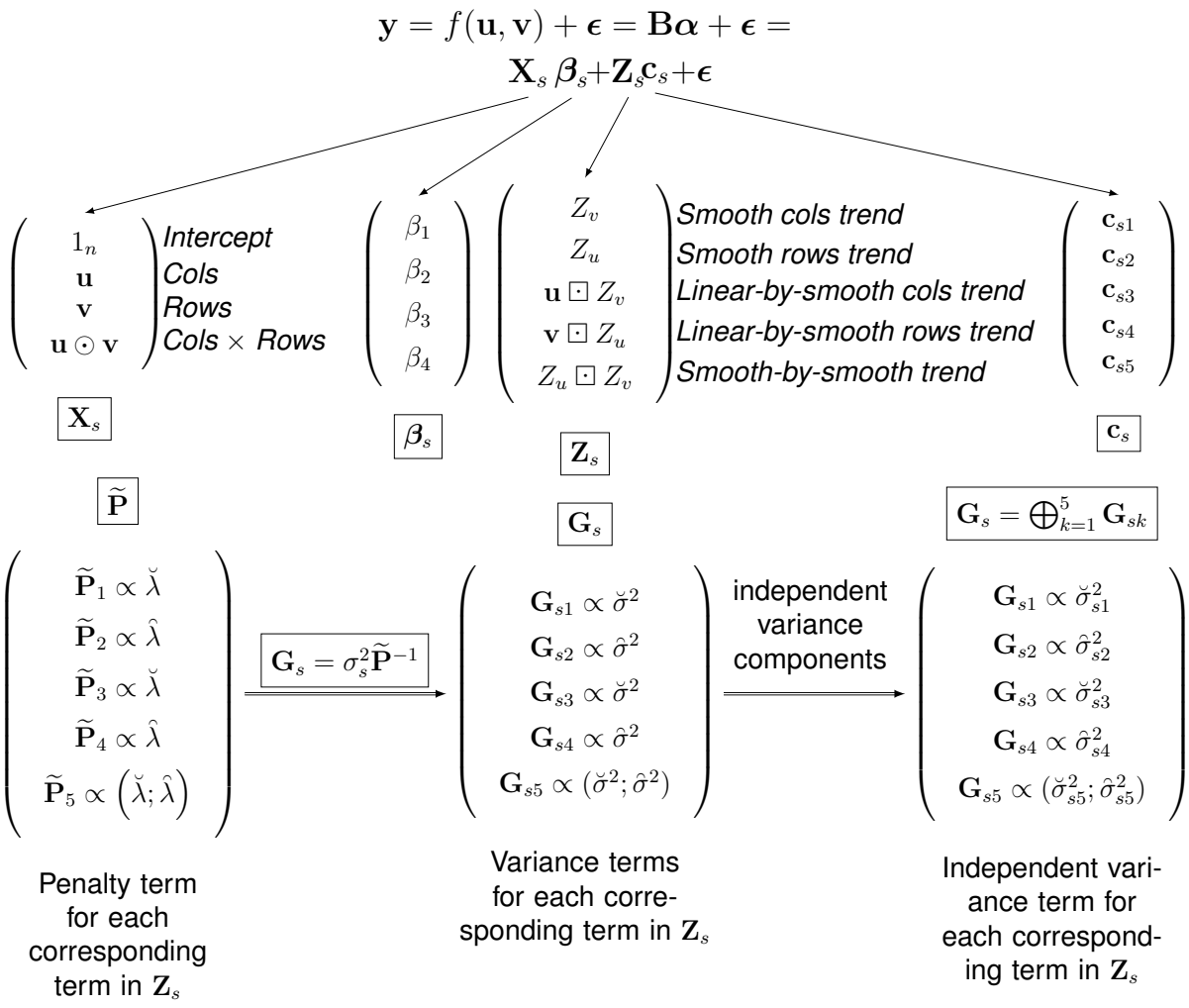


Figure 3.7: Diagram detailing the structure of the matrices used in this section. All matrices are block diagonal matrix with each element represented on the diagram, being an individual block. The symbol $\propto$ shows how each block of the $\widetilde{\mathbf{P}}/\mathbf{G}_s$ matrix relates to the tuning/variance parameters. The last block of both the $\widetilde{\mathbf{P}}$ and $\mathbf{G}_s$ matrices depends on both parameters but in a non-linear way.

3.4.4 Model estimation

With all these specifications in mind, the model was fitted using cubic B-splines and second-order penalties. These settings are commonly used and allow flexibility of the model (M. X. Rodríguez-Álvarez, Boer, van Eeuwijk & P. H. C. Eilers 2016, 2018; M. X. Rodríguez-Álvarez, Lee, *et al.* 2015). We used 99 and 5 equally spaced knots for the the P-splines, corresponding to the strips and positions, respectively. In this way there was approximately one knot for every plot. Then there was a total of 362 model parameters to be estimated for the smooth surface. As M. X. Rodríguez-Álvarez, Boer, van Eeuwijk & P. H. Eilers (2018) explains, the number of knot is not critical since the optimization of the fit to the data is essentially dependant on the smoothing parameters. The estimation procedure was performed using the R-package `SpATS` (M. Rodríguez-Álvarez *et al.* 2016). This package provides a REML-based estimation of the variances components and computes the best linear unbiased estimators (BLUEs) of the fixed effects and the empirical best linear unbiased predictors (BLUPs) of the random effects. A useful by-product of this computation is the effective dimension associated to each random effect.

Effective dimensions

In P-splines methodology, the effective dimension (ED) measures the complexity of the model components (P. H. Eilers *et al.* 2015), it is similar to the more common concept of effective degree of freedom (Buja *et al.* 1989). It is computed as the trace of the hat matrix $\mathbf{H}$. If we only take the spatial part of our model (equation 3.4.24), we can write:

$$\begin{aligned} \tilde{\mathbf{y}} &= \mathbf{H}\mathbf{y} \\ \tilde{f}(\mathbf{u},\mathbf{v}) &= \boldsymbol{X}_s\hat{\boldsymbol{\beta}}_s + \boldsymbol{Z}_s\tilde{\boldsymbol{c}}_s \ , \\ &= \mathbf{H}_\beta\mathbf{y} + \mathbf{H}_s\mathbf{y} \end{aligned} \tag{3.4.41}$$

where $\mathbf{H}_\beta$ is the hat matrix of the fixed components and $\mathbf{H}_s$ is the hat matrix of the random components, also known as the smoother matrix. In this context, the sum of the diagonal elements of $\mathbf{H}_s$ expresses the number of parameters effectively involved in the modelling of the spatial surface. This decomposition is allowed from the PS-ANOVA structure of the spatial model. Following this, the smoother matrix can be further decomposed according to the five additive and interaction smooth components of the smooth bivariate surface, giving $\mathbf{H}_s = \sum_{k=1}^{5}\mathbf{H}_{s_k}$, and we can compute the individual effective dimension for each component (ED_{s_k}). As explained in M. X. Rodríguez-Álvarez, Boer, van Eeuwijk & P. H. C. Eilers (2016), the effective dimension varies with the smoothing parameter:

$$\begin{aligned} \lambda_{s_k} = \frac{\sigma^2_e}{\sigma^2_{s_k}} \to \infty \text{ then } ED_{s_k} \to 0 \\ \lambda_{s_k} = \frac{\sigma^2_e}{\sigma^2_{s_k}} \to 0 \text{ then } ED_{s_k} \to \text{upper bound} \ , \end{aligned} \tag{3.4.42}$$

where the upper bound is determined by the number of knots.

Consequently, the total effective dimension ED_s can be interpreted as a measure of the magnitude of field variations with larger values indicating more intense spatial patterns. In addition, the partial effective dimensions ED_{s_k} are indicative of the contribution of each component to the fitted surface, and reflect the complexity of the spatial pattern.

3.5 ARXAR model

In this section the ARXAR model, and its extension to the linear variance (LV) model, are explained. For more detailed information about the original ARxAR model, consult Gilmour *et al.* (1997). For information about the extensions of the model, see **Piepho2010** and E. R. Williams (1986).
Let us consider a similar starting point as for the SpATS model, a field trial of n plots arranged in a rectangular grid, where the plot positions are collected in vectors of row $(\mathbf{r})$ and column $(\mathbf{c})$ coordinates, and $\mathbf{y}$ is the vector of plot data in field order. Here the starting model for $\mathbf{y}$ is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{c} + \boldsymbol{\xi} + \boldsymbol{\eta} \tag{3.5.01}$$

where $\boldsymbol{\beta}^{(t\times 1)}$ is the vector of fixed effects with design matrix $\mathbf{X}^{(n\times t)}$, $\mathbf{c}^{(b\times 1)}$ is the vector of random effects with design matrix $\mathbf{Z}^{(n\times b)}$, $\boldsymbol{\xi}^{(n\times 1)}$ is a spatially dependent random error vector and $\boldsymbol{\eta}^{(n\times 1)}$ is a zero mean random error vector whose elements are pairwise independent. It is also assumed that $(\mathbf{c}, \boldsymbol{\xi}, \boldsymbol{\eta})$ are pairwise independent and that their joint distribution is Gaussian with zero mean and variance

$$\sigma^2 \begin{bmatrix} \mathbf{G}(\boldsymbol{\gamma}) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \Sigma(\boldsymbol{\alpha}) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \psi\mathbf{I} \end{bmatrix}, \tag{3.5.02}$$

where $\psi = \sigma^2_\eta/\sigma^2$, $\boldsymbol{\gamma}$ is the vector of variance components ratios corresponding to possible subvectors in $\mathbf{c}$, and $\boldsymbol{\alpha}$ is a vector of spatial covariance parameters. The marginal distribution of $\mathbf{y}$ is then

$$\mathbf{y} \sim \mathbf{N}\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2\left(\mathbf{Z}\mathbf{G}\mathbf{Z}^t + \mathbf{R}\right)\right), \tag{3.5.03}$$

where $\mathbf{R} = \mathbf{R}(\phi) = \Sigma + \psi\mathbf{I}, \phi = (\boldsymbol{\alpha}^t, \psi)^t$. As said in previous sections, the goal of the ARXAR model is to use a variogram to estimate the errors terms of model 3.5.01 and therefore produce better estimates of the fixed effects of the model.

3.5.1 The variogram

Given a spatially correlated error process $\mathcal{E}(.)$ at point $\mathbf{s}$ and $\mathbf{t}$, the theoretical variogram (also called the semi-variogram) of $\mathcal{E}(.)$ is the function

$$\omega(\mathbf{s}, \mathbf{t}) = \frac{1}{2}\operatorname{var}[\mathcal{E}(\mathbf{s}) - \mathcal{E}(\mathbf{t})] = \frac{1}{2}[\mathrm{V}(\mathbf{s}, \mathbf{s}) + \mathrm{V}(\mathbf{t}, \mathbf{t}) - 2\mathrm{V}(\mathbf{s}, \mathbf{t})], \tag{3.5.11}$$

where $\mathbf{s}, \mathbf{t} \in \Re^2$ and $\mathbf{V}(.,.)$ is the covariance function of $\mathcal{E}(.)$. Here, $\mathcal{E}(.)$ is assumed to be second-order stationary. To illustrate these concepts, we consider $\mathbf{e} = \boldsymbol{\xi} + \boldsymbol{\eta}$ where

$\mathbf{e}$ is a zero mean spatially correlated process with a directional exponential covariance (DEC) structure distributed independently of $\boldsymbol{\eta}$, which is a zero mean white-noise process (**cressie1992statistics**). Let

$$\mathbf{l} = \begin{bmatrix} l_1 \\ l_2 \end{bmatrix} = \begin{bmatrix} |s_1 - t_1| \\ |s_2 - t_2| \end{bmatrix}$$

be the "distance" between points $\mathbf{s}$ and $\mathbf{t}$. Then

$$\begin{aligned} \omega(\mathbf{s}, \mathbf{t}) = \omega(\mathbf{l}) \quad &= \sigma_\eta^2 + \sigma^2 \left[1 - \exp\left(-\alpha_1 l_1 - \alpha_2 l_2\right)\right] \quad \mathbf{l} \neq 0 \\ &= 0 \qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad \mathbf{l} = 0 \end{aligned} . \tag{3.5.12}$$

The measurement error term induces a jump discontinuity at $\mathbf{l} = 0$. For most field experiments, where plots are arranged in regular arrays and therefore separated by equivalent distances, the displacement vector takes values for l_1 of $0, d_1, 2d_1, \ldots, (r-1)d_1$ and for l_2 of $0, d_2, 2d_2, \ldots, (c-1)d_2$, where d_1 and d_2 are the plot dimensions. Then the previous equation can be rewritten as a function of an indexed displacement vector $\mathbf{l}^*$ with values for l_1^* of $0, 1, 2, \ldots, (r-1)$ and values for l_2^* of $0, 1, 2, \ldots, (c-1)$, and becomes

$$\begin{aligned} \omega\left(\mathbf{l}^*\right) &= \sigma_\eta^2 + \sigma^2 \left[1 - \exp\left(-\alpha_1 d_1 l_1^* - \alpha_2 d_2 l_2^*\right)\right] \\ &= \sigma_\eta^2 + \sigma^2 \left(1 - \rho_1^{l_1^*} \rho_2^{l_2^*}\right) \qquad \mathbf{l}^* \neq 0 \\ &= 0 \qquad\qquad\qquad\qquad\qquad\qquad\quad \mathbf{l}^* = 0 \end{aligned} \tag{3.5.13}$$

where $\rho_1 = \exp\left(-\alpha_1 d_1\right)$ and $\rho_2 = \exp\left(-\alpha_2 d_2\right)$. This formulation demonstrates the equivalence between the DEC model and the AR1 x AR1 model for field experiments. Considering the model given in equation 3.5.01, the variogram ordinates for the data vector $\mathbf{y}$ are

$$v_{ij} = \frac{1}{2}\left[e_i\left(\mathbf{s}_i\right) - e_j\left(\mathbf{s}_j\right)\right]^2 \quad \forall i, j = 1, \ldots, n; i \neq j \tag{3.5.14}$$

where $\mathbf{e} = \{e_i(\mathbf{s}_i)\} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{c}$. When $\boldsymbol{\beta}$ and $\mathbf{c}$ are known and under the assumption that $\mathbf{y}$ is Gaussian, the sampling distribution of v_{ij} is

$$\frac{v_{ij}}{\omega\left(\mathbf{s}_i, \mathbf{s}_j\right)} \sim \chi_1^2 \tag{3.5.15}$$

so that v_{ij} is unbiased for $\omega\left(\mathbf{s}_i, \mathbf{s}_j\right)$. As implied previously, many v_{ij} will have the same absolute displacement since the plots are arranged in a regular array. Therefore the sample variogram is presented as the triplet $(l_{ij1}, l_{ij2}, \overline{v}_{ij})$, where $l_{ij1} = |s_{i1} - s_{j1}|$ and $l_{ij2} = |s_{i2} - s_{j2}|$ are the absolute displacements and $\overline{v}_{ij}$ is the sample mean of the v_{ij} with the same absolute displacements.

3.5.2 Model estimation

The result that v_{ij} is unbiased for $\omega\left(\mathbf{s}_i, \mathbf{s}_j\right)$ is based on the assumptions that $\boldsymbol{\beta}$ and $\mathbf{c}$ are known. In practice, we replace $\boldsymbol{\beta}$ and $\mathbf{c}$ by their GLS estimates $(\widehat{\boldsymbol{\beta}})$ and BLUP $(\hat{\mathbf{c}})$ respectively, so that the BLUP of the residual vector is given by

$$\hat{\mathbf{e}} \quad = \quad \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{Z}\widehat{\mathbf{c}} = \mathbf{R}\mathbf{P}\mathbf{y} \tag{3.5.21}$$

where $\mathbf{P} = \mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{W}\mathbf{C}^{-1}\mathbf{W}^t\mathbf{R}^{-1}$ with $\boldsymbol{W} = [\boldsymbol{X}\mathbf{Z}]$ and $\mathbf{C} = \mathbf{W}^t\mathbf{R}^{-1}\mathbf{W} + \mathbf{G}^*$ is the coefficient matrix from the mixed model equation and it partitioned in the same way as $\mathbf{W}$. $\mathbf{G}^*$ is a square matrix of order $t+b$, partitioned similarly to $\mathbf{W}^t\mathbf{R}^{-1}\mathbf{W}$ and is 0 expect in the lower diagonal block corresponding to $\mathbf{Z}^t\mathbf{R}^{-1}\mathbf{Z}$, where it equals $\mathbf{G}^{-1}$.
Under the assumption of a Gaussian distribution for $\mathbf{y}$, $\hat{\mathbf{e}} \sim \mathrm{N}\left(\mathbf{0}, \sigma^2\left(\mathbf{R} - \mathbf{W}\mathbf{C}^{-1}\mathbf{W}^t\right)\right)$ assuming (γ, ϕ) is known. Following the decomposition of (3.5.14), variogram ordinates v_{ij} can be expressed as a quadratic form in $\mathbf{y}$, that is

$$v_{ij} = \left(\mathbf{a}_{ij}^t\mathbf{e}\right)\left(\mathbf{a}_{ij}^t\mathbf{e}\right) = \mathbf{e}^t\mathbf{a}_{ij}\mathbf{a}_{ij}^t\mathbf{e} = \mathbf{e}^t\mathbf{A}_{ij}\mathbf{e} \tag{3.5.22}$$

and similarly

$$\hat{v}_{ij} = \left(\mathbf{a}_{ij}^t\hat{\mathbf{e}}\right)\left(\mathbf{a}_{ij}^t\hat{\mathbf{e}}\right) = \hat{\mathbf{e}}^t\mathbf{a}_{ij}\mathbf{a}_{ij}^t\hat{\mathbf{e}} = \hat{\mathbf{e}}^t\mathbf{A}_{ij}\hat{\mathbf{e}} \tag{3.5.23}$$

where $\mathbf{A}_{ij}^{(n\times n)}$ has a $1/2$ value in positions $\{i,i\}$ and $\{j,j\}$, a $-1/2$ value in positions $\{i,j\}$ and $\{j,i\}$ and 0 elsewhere.
Taking the expectation

$$\begin{aligned}
\mathrm{E}\left(\hat{v}_{ij}\right) &= \sigma^2 \operatorname{trace}\left[\mathbf{A}_{ij}\left(\mathbf{R} - \mathbf{W}\mathbf{C}^{-1}\mathbf{W}^t\right)\right] \\
&= \sigma^2 \operatorname{trace}\left[\mathbf{A}_{ij}\mathbf{R}\right] - \sigma^2 \text{ trace } \left[\mathbf{A}_{ij}\mathbf{W}\mathbf{C}^{-1}\mathbf{W}^t\right] \\
&= \sigma^2\mathbf{a}_{ij}^t\mathbf{R}\mathbf{a}_{ij} - \sigma^2\mathbf{a}_{ij}^t\mathbf{W}\mathbf{C}^{-1}\mathbf{W}^t\mathbf{a}_{ij} \\
&= \quad \mathrm{E}\left(v_{ij}\right) - \sigma^2\mathbf{a}_{ij}^t\mathbf{W}\mathbf{C}^{-1}\mathbf{W}^t\mathbf{a}_{ij}
\end{aligned} \tag{3.5.24}$$

Thus $\tilde{v}_{ij}$ is biased. However the bias can be removed by considering the spectral decomposition of $\mathbf{Z}\boldsymbol{G}\mathbf{Z}^t$ which has $t+b$ non-zero eigenvalues. Let

$$\mathbf{W}\mathbf{C}^{-1}\mathbf{W}' = \sum_{k=1}^{t+b}\lambda_k\mathbf{w}_k\mathbf{w}_k', \tag{3.5.25}$$

then

$$\begin{aligned}
\mathrm{E}\left(\hat{v}_{ij}\right) &= \mathrm{E}\left(v_{ij}\right) - \sigma^2\mathbf{a}_{ij}'\left(\sum_k \lambda_k\mathbf{w}_k\mathbf{w}_k'\right)\mathbf{a}_{ij} \\
&= \mathrm{E}\left(v_{ij}\right) - \sigma^2\sum_h \lambda_k\left(\mathbf{a}_{ij}'\mathbf{w}_k\right)^2 \\
&= \mathrm{E}\left(v_{ij}\right) - \sigma^2\sum_k \lambda_k\mathbf{w}_k'\mathbf{A}_{ij}\mathbf{w}_k
\end{aligned} \tag{3.5.26}$$

Thus, the bias in $\tilde{v}_{ij}$ is easily calculated as the weighted sum of the variogram ordinates for each of the $t+b$ eigenvectors $\mathbf{w}_k$. In practice, we are concerned with the general shape of the variogram, so it is often sufficient to use only the largest r eigenvalues and their corresponding eigenvectors, where r is much smaller than $t+b$. This derivation assumes (γ, ϕ) are known. In practice these are replaced by their REML estimates, so 3.5.26 is approximate. The effect of the estimation of (γ, ϕ) on the distribution of $\hat{\mathbf{e}}$ (and functions of $\hat{\mathbf{e}}$) is an important problem. **kenward1997precision** have examined this issue for the testing of fixed effects in REML.

3.5.3 Base model selection

"Do a section explaining model selection and which one was retained for our dataset. " See velazco ang gilmour for selection.

3.5.4 Extension to the linear variance (LV) model

Explain how the model were fitted

3.6 Model comparison

The SpATS model was compared with the BSS models in terms of meaningful parameters for plant breeding application. The estimates considered for comparison are similar to those used by Velazco *et al.* (2017). The following estimates are:

- Genetic variance (σ_g^2) and spatially independent residual variance (σ_e^2). The goal being having a minimal variance in both cases.
- Generalized heritability. Estimated as described previously for the SpATS model and as described by Cullis, Smith, *et al.* (2006) for the BSS model. In this case, the heritability is interpreted as the measure of the precision of a trial, i.e. the ability to detect genotypic differences among test-cross means. Given that our study does not incorporate a genetic relationship matrix, we can use equation **??** to perform a straightforward comparison between the heritability estimated by SpATS and that obtained from the BSS model (Velazco *et al.* 2017).
- Spearman rank correlation between predicted genotypic values for the different models in the same environment. This will allow us to compare the ranking of genotypes from the SpATS model and from the BSS model.

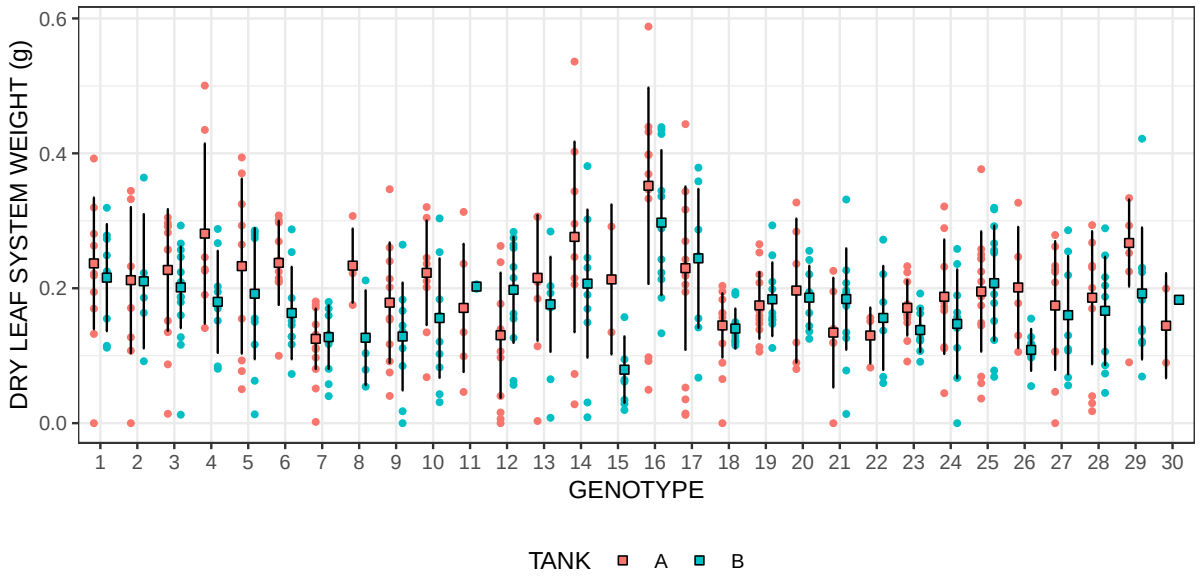
It is interesting to note that M. X. Rodríguez-Álvarez, Boer, van Eeuwijk & P. H. Eilers (2018) also use the Pearson correlations of predicted genotypes values between environments (i.e. field trials) as a way of comparing models. Since only one trial was studied in this thesis, this correlation cannot be used.

Chapter 4

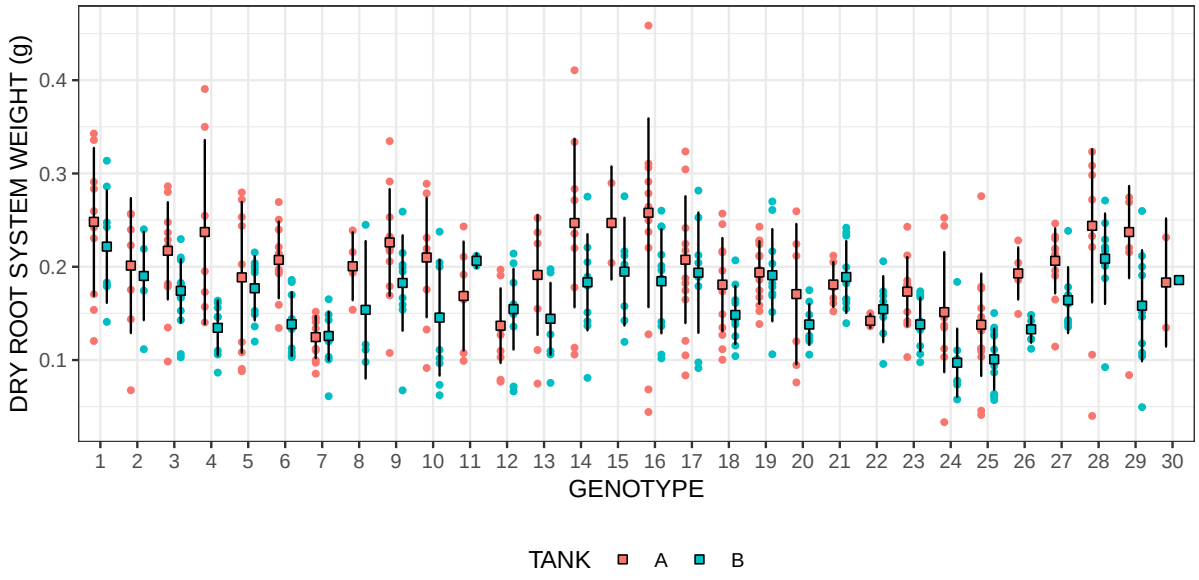
Results and discussion

4.1 Pre-processing

Before the experiment, the seeds were weighted in groups of ten, to see if there was a baseline difference between certain genotypes. Table 3.1, in the previous section, displays the measured weights. After the completion of the experiment, the outliers were identified, and each plant was attributed a specific weight, following the protocol described in material and methods section. Those weights were used to compute the weighted mean and weighted standard deviation of the fresh and dry weight of the root system and the leaf system, as well as the area of the root system, for each plant. These results are presented in table 4.1.

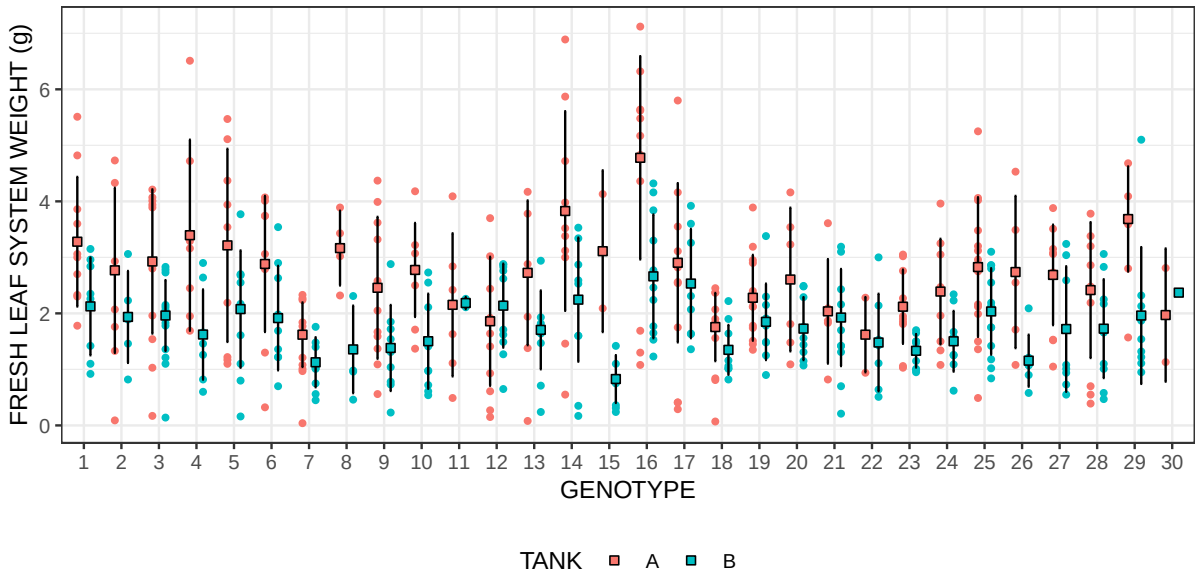


(a) Dry leaf weight (DRY_LS

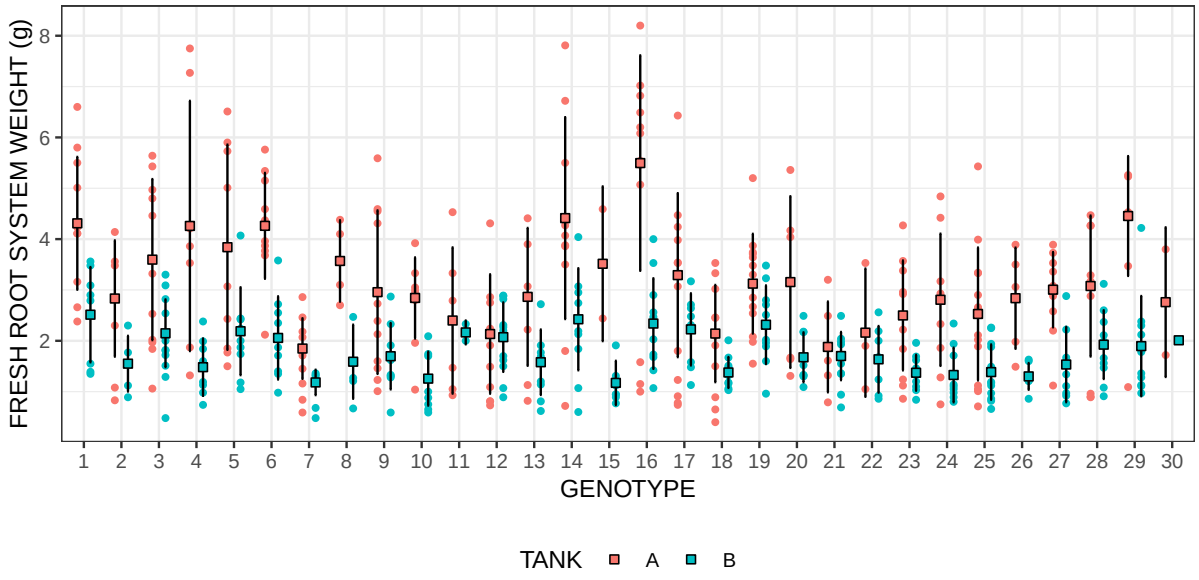


(b) Dry root weight (DRY_RS)

Figure 4.1: Dotplot displaying mean weight (square) and associated standard deviation (black line), grouped by tanks for each variable.



(c) Fresh leaf weight ($FRESH\_LS$)



(d) Fresh root weight ($FRESH\_RS$)

Figure 4.1: Dotplot displaying mean weight (□) and associated standard deviation (—), grouped by tanks for each variable.

Table 4.1: Weighted mean and standard deviation for each genotype. $\mathbf{DRY_{LS}}$ represents the dry weight of the leaf system; $\mathbf{DRY_{RS}}$, the dry weight for the root system; $\mathbf{FRESH_{LS}}$, the fresh weight for the leaf system and $\mathbf{FRESH_{RS}}$, the fresh weight for the root system. All the results are presented as mean $\pm$ standard deviation (g)

Genotype	DRY_{LS}	DRY_{RS}	$FRESH_{LS}$	$FRESH_{RS}$
1	0.2267 $\pm$ 0.0869	0.2354 $\pm$ 0.0698	2.7231 $\pm$ 1.1612	3.4447 $\pm$ 1.4431
2	0.2113 $\pm$ 0.0993	0.1964 $\pm$ 0.06	2.4058 $\pm$ 1.254	2.2725 $\pm$ 1.1119
3	0.2132 $\pm$ 0.0747	0.1939 $\pm$ 0.0474	2.406 $\pm$ 1.0814	2.8146 $\pm$ 1.3663
4	0.227 $\pm$ 0.1148	0.1824 $\pm$ 0.0861	2.45 $\pm$ 1.5508	2.7773 $\pm$ 2.1926
5	0.2126 $\pm$ 0.113	0.1829 $\pm$ 0.0606	2.6521 $\pm$ 1.5044	3.0241 $\pm$ 1.7336
6	0.2024 $\pm$ 0.0739	0.1747 $\pm$ 0.051	2.4244 $\pm$ 1.1691	3.2168 $\pm$ 1.4545
7	0.126 $\pm$ 0.0441	0.1251 $\pm$ 0.0232	1.4118 $\pm$ 0.5677	1.5669 $\pm$ 0.5783
8	0.186 $\pm$ 0.0805	0.1798 $\pm$ 0.0567	2.3614 $\pm$ 1.1696	2.6894 $\pm$ 1.273
9	0.1559 $\pm$ 0.0865	0.2064 $\pm$ 0.0574	1.9704 $\pm$ 1.1782	2.3862 $\pm$ 1.3949
10	0.1885 $\pm$ 0.0875	0.1769 $\pm$ 0.0693	2.1228 $\pm$ 1.046	2.029 $\pm$ 1.0451
11	0.1789 $\pm$ 0.0811	0.1783 $\pm$ 0.0521	2.1608 $\pm$ 1.0747	2.339 $\pm$ 1.2172
12	0.1684 $\pm$ 0.0893	0.1469 $\pm$ 0.0417	2.0166 $\pm$ 0.9281	2.0985 $\pm$ 0.9015
13	0.1927 $\pm$ 0.0794	0.1639 $\pm$ 0.054	2.1326 $\pm$ 1.077	2.1176 $\pm$ 1.1582
14	0.2438 $\pm$ 0.1288	0.2173 $\pm$ 0.0796	3.0901 $\pm$ 1.6735	3.486 $\pm$ 1.8646
15	0.1175 $\pm$ 0.0885	0.2097 $\pm$ 0.0591	1.48 $\pm$ 1.2962	1.8417 $\pm$ 1.3372
16	0.3244 $\pm$ 0.1276	0.2208 $\pm$ 0.0879	3.7094 $\pm$ 1.8196	3.9028 $\pm$ 2.2598
17	0.2357 $\pm$ 0.1111	0.2019 $\pm$ 0.065	2.7519 $\pm$ 1.2418	2.8555 $\pm$ 1.4
18	0.1427 $\pm$ 0.039	0.1653 $\pm$ 0.0442	1.5621 $\pm$ 0.5627	1.7781 $\pm$ 0.8075
19	0.1785 $\pm$ 0.0505	0.1926 $\pm$ 0.04	2.0917 $\pm$ 0.7454	2.7726 $\pm$ 0.9668
20	0.1901 $\pm$ 0.0721	0.1506 $\pm$ 0.0497	2.0641 $\pm$ 0.9693	2.2429 $\pm$ 1.2908
21	0.1649 $\pm$ 0.0786	0.1859 $\pm$ 0.0329	1.9682 $\pm$ 0.8601	1.7689 $\pm$ 0.6407
22	0.1482 $\pm$ 0.0669	0.1508 $\pm$ 0.0296	1.522 $\pm$ 0.7799	1.793 $\pm$ 0.838
23	0.156 $\pm$ 0.04	0.1574 $\pm$ 0.0372	1.7621 $\pm$ 0.6537	1.9874 $\pm$ 0.9962
24	0.1668 $\pm$ 0.0827	0.1236 $\pm$ 0.0575	1.9357 $\pm$ 0.8693	2.0506 $\pm$ 1.2218
25	0.2013 $\pm$ 0.0856	0.1195 $\pm$ 0.0481	2.4338 $\pm$ 1.0937	1.9605 $\pm$ 1.1428
26	0.1463 $\pm$ 0.075	0.1573 $\pm$ 0.0365	1.7995 $\pm$ 1.1967	1.9256 $\pm$ 1.0053
27	0.1682 $\pm$ 0.0894	0.188 $\pm$ 0.0401	2.2662 $\pm$ 1.0855	2.3647 $\pm$ 1.0382
28	0.1753 $\pm$ 0.0865	0.2244 $\pm$ 0.066	2.0357 $\pm$ 1.0699	2.441 $\pm$ 1.1785
29	0.215 $\pm$ 0.0934	0.1823 $\pm$ 0.0663	2.4861 $\pm$ 1.3776	2.6743 $\pm$ 1.5787
30	0.1573 $\pm$ 0.0592	0.184 $\pm$ 0.0485	2.1033 $\pm$ 0.8712	2.51 $\pm$ 1.1265

4.2 SpATS analysis

4.3 ARxAR model analysis

4.4 Model comparison

4.4.1 Performances

4.4.2 Parametrization

4.4.3 Modelling strategy

Chapter 5

Conclusion

Bibliography

1. Atkinson, A. C. Optimal Design. *Wiley StatsRef: Statistics Reference Online,* 1–17 (2014).
2. Atkinson, A. C. & Bailey, R. A. One hundred years of the design of experiments on and off the pages of Biometrika. en. *Biometrika* **88,** 53–97 (Feb. 2001).
3. Atkinson, A. C. & Donev, A. N. The construction of exact D-optimum experimental designs with application to blocking response surface designs. *Biometrika* **76,** 515–526 (1989).
4. Bohachevsky, I. O., Johnson, M. E. & Stein, M. L. Generalized simulated annealing for function optimization. *Technometrics* **28,** 209–217 (1986).
5. Brien, C. J., Berger, B., Rabie, H. & Tester, M. Accounting for variation in designing greenhouse experiments with special reference to greenhouses containing plants on conveyor systems. *Plant Methods* **9,** 5 (Feb. 2013).
6. Buja, A., Hastie, T., Tibshirani, R., *et al.* Linear smoothers and additive models. *The Annals of Statistics* **17,** 453–510 (1989).
7. Cabrera-Bosquet, L. *et al.* High-throughput estimation of incident light, light interception and radiation-use efficiency of thousands of plants in a phenotyping platform. en. *New Phytologist* **212,** 269–281 (Oct. 2016).
8. Cullis, B. R. & Gleeson, A. C. Spatial Analysis of Field Experiments-An Extension to Two Dimensions. *Biometrics* **47,** 1449–1460 (1991).
9. Cullis, B. R., Smith, A. B. & Coombes, N. E. On the design of early generation variety trials with correlated data. en. *Journal of Agricultural, Biological, and Environmental Statistics* **11,** 381 (Dec. 2006).
10. Currie, I. D. & Durban, M. Flexible smoothing with P-splines: a unified approach. en. *Statistical Modelling* **2,** 333–349 (Dec. 2002).
11. Davidoff, B., Lewis, J. W. & Selim, H. M. A method to verify the presence of a trend in studying spatial variability of soil temperature. English. *Soil Science Society of America journal (USA)* (1986).
12. Dierckx, P. *Curve and Surface Fitting with Splines* en (Clarendon Press, 1995).
13. Durban, M., Currie, I. D. & Kempton, R. A. Adjusting for fertility and competition in variety trials. en. *The Journal of Agricultural Science* **136,** 129–140 (Mar. 2001).
14. Eilers, P. H. C. & Marx, B. D. Flexible smoothing with B-splines and penalties. en. *Statistical Science* **11,** 89–121 (May 1996).

15. Eilers, P. H. C. & Marx, B. D. Multivariate calibration with temperature interaction using two-dimensional penalized signal regression. *Chemometrics and Intelligent Laboratory Systems* **66,** 159–174 (June 2003).

16. Eilers, P. H., Marx, B. D. & Durbán, M. Twenty years of P-splines. *SORT: statistics and operations research transactions* **39,** 0149–186 (2015).

17. Fagroud, M. & Van Meirvenne, M. Accounting for Soil Spatial Autocorrelation in the Design of Experimental Trials. en. *Soil Science Society of America Journal* **66,** 1134–1142 (July 2002).

18. Fahrmeir, L., Kneib, T., Lang, S. & Marx, B. *Regression: Models, Methods and Applications* en (Springer-Verlag, Berlin Heidelberg, 2013).

19. Fedorov, V. V. *Theory of optimal experiments* eng. Open Library ID: OL18496755M (Academic Press, New York, 1972).

20. Fiorani, F. & Schurr, U. Future Scenarios for Plant Phenotyping. en. *Annual Review of Plant Biology* **64,** 267–291 (Apr. 2013).

21. Furbank, R. T. & Tester, M. Phenomics – technologies to relieve the phenotyping bottleneck. en. *Trends in Plant Science* **16,** 635–644 (Dec. 2011).

22. Gilmour, A. R., Cullis, B. R. & Verbyla, A. P. Accounting for Natural and Extraneous Variation in the Analysis of Field Experiments. *Journal of Agricultural, Biological, and Environmental Statistics* **2,** 269–293 (1997).

23. Goos, P. & Jones, B. *Optimal Design of Experiments: A Case Study Approach* en. Google-Books-ID: EMWYkYd3sPoC (John Wiley & Sons, June 2011).

24. Heredia-Langner, A., Carlyle, W. M., Montgomery, D. C., Borror, C. M. & Runger, G. C. Genetic algorithms for the construction of D-optimal designs. *Journal of Quality Technology* **35,** 28–46 (2003).

25. Heredia-Langner, A., Montgomery, D. C., Carlyle, W. M. & Borror, C. M. Model-robust optimal designs: A genetic algorithm approach. *Journal of Quality Technology* **36,** 263–279 (2004).

26. Houle, D., Govindaraju, D. R. & Omholt, S. Phenomics: the next challenge. en. *Nature Reviews Genetics* **11,** 855–866 (Dec. 2010).

27. Iqbal, J., Thomasson, J. A., Jenkins, J. N., Owens, P. R. & Whisler, F. D. Spatial Variability Analysis of Soil Physical Properties of Alluvial Soils. en. *Soil Science Society of America Journal* **69,** 1338 (2005).

28. Johnson, M. E. & Nachtsheim, C. J. Some guidelines for constructing exact D-optimal designs on convex design spaces. *Technometrics* **25,** 271–277 (1983).

29. Jung, J. S. & Yum, B. J. Construction of exact D-optimal designs by tabu search. *Computational Statistics & Data Analysis* **21,** 181–191 (1996).

30. Lado, B. *et al.* Increased Genomic Prediction Accuracy in Wheat Breeding Through Spatial Adjustment of Field Trial Data. en. *G3: Genes, Genomes, Genetics* **3,** 2105–2114 (Dec. 2013).

31. Lee Hwang, D.-J. Smoothing mixed models for spatial and spatio-temporal data. eng (May 2010).

32. Lee, D.-J. & Durbán, M. P-spline ANOVA-type interaction models for spatio-temporal smoothing. *Statistical Modelling* **11,** 49–69 (Feb. 2011).

33. Lee, D.-J., Durbán, M. & Eilers, P. Efficient two-dimensional smoothing with P-spline ANOVA mixed models and nested bases. *Computational Statistics & Data Analysis* **61,** 22–37 (May 2013).

34. Lobet, G. & Draye, X. Novel scanning procedure enabling the vectorization of entire rhizotron-grown root systems. en. *Plant Methods* **9,** 1 (2013).

35. Lobet, G., Draye, X. & Périlleux, C. An online database for plant image analysis software tools. *Plant Methods* **9,** 38 (Oct. 2013).

36. Lobet, G., Pagès, L. & Draye, X. A Novel Image-Analysis Toolbox Enabling Quantitative Analysis of Root System Architecture. en. *Plant Physiology* **157,** 29–39 (Sept. 2011).

37. Meyer, R. K. & Nachtsheim, C. J. Constructing Exact D-Optimal Experimental Designs by Simulated Annealing. *American Journal of Mathematical and Management Sciences* **8,** 329–359 (Feb. 1988).

38. Meyer, R. K. & Nachtsheim, C. J. The coordinate-exchange algorithm for constructing exact optimal experimental designs. English (US). *Technometrics* **37,** 60–69 (Jan. 1995).

39. Mooney, S. J., Pridmore, T. P., Helliwell, J. & Bennett, M. J. Developing X-ray Computed Tomography to non-invasively image 3-D root systems architecture in soil. en. *Plant and Soil* **352,** 1–22 (Mar. 2012).

40. Nielsen, D., Biggar, J. & Erh, K. Spatial variability of field-measured soil-water properties. English. *Hilgardia* **42,** 215–259 (Nov. 1973).

41. Oakey, H., Verbyla, A., Pitchford, W., Cullis, B. & Kuchel, H. Joint modeling of additive and non-additive genetic line effects in single field trials. en. *Theoretical and Applied Genetics* **113,** 809–819 (Sept. 2006).

42. Piepho, H. P. & Williams, E. R. Linear variance models for plant breeding trials. en. *Plant Breeding* **129,** 1–8 (Feb. 2010).

43. Piepho, H., Möhring, J., Pflugfelder, M., Hermann, W. & Williams, E. Problems in parameter estimation for power and AR(1) models of spatial correlation in designed field experiments. *Communications in Biometry and Crop Science* **10,** 3–16 (2015).

44. Pieruschka, R., Schurr, U., *et al.* Plant Phenotyping: Past, Present, and Future. *Plant Phenomics* **2019,** 7507131 (2019).

45. Pound, M. P. *et al.* RootNav: Navigating Images of Complex Root Architectures. en. *PLANT PHYSIOLOGY* **162,** 1802–1814 (Aug. 2013).

46. Risser, M. D. Nonstationary Spatial Modeling, with Emphasis on Process Convolution and Covariate-Driven Approaches. *arXiv preprint arXiv:1610.02447* (2016).

47. Rodríguez-Álvarez, M. X., Boer, M. P., van Eeuwijk, F. A. & Eilers, P. H. C. Spatial Models for Field Trials. en. *arXiv:1607.08255 [stat].* arXiv: 1607.08255 (July 2016).

48. Rodríguez-Álvarez, M. X., Boer, M. P., van Eeuwijk, F. A. & Eilers, P. H. Correcting for spatial heterogeneity in plant breeding experiments with P-splines. en. *Spatial Statistics* **23,** 52–71 (Mar. 2018).

49. Rodríguez-Álvarez, M. X., Lee, D.-J., Kneib, T., Durbán, M. & Eilers, P. Fast smoothing parameter separation in multidimensional generalized P-splines: the SAP algorithm. *Statistics and Computing* **25,** 941–957 (2015).

50. Rodríguez-Álvarez, M., Boer, M., Eilers, P. & van Eeuwijk, F. *SpATS: spatial analysis of field trials with splines. R package version 1.0–4* (2016).

51. Rodríguez, M., Jones, B., Borror, C. M. & Montgomery, D. C. Generating and Assessing Exact G-Optimal Designs. *Journal of Quality Technology* **42,** 3–20 (Jan. 2010).

52. Tardieu, F., Cabrera-Bosquet, L., Pridmore, T. & Bennett, M. Plant Phenomics, From Sensors to Knowledge. en. *Current Biology* **27,** R770–R783 (Aug. 2017).

53. Van Es, H. M. 1.2 Soil Variability. *Methods of Soil Analysis: Part 4 Physical Methods,* 1–13 (2002).

54. Van Es, H. M. Spatial Nature of Randomization and Its Effect on the Outcome of Field Experiments. en. *Agronomy Journal* **85,** 420–428 (1993).

55. Van Es, H. M., Gomes, C. P., Sellmann, M. & van Es, C. L. Spatially-Balanced Complete Block designs for field experiments. *Geoderma. Pedometrics 2005* **140,** 346–352 (Aug. 2007).

56. Vargas, M., van Eeuwijk, F. A., Crossa, J. & Ribaut, J.-M. Mapping QTLs and QTL × environment interaction for CIMMYT maize drought stress program using factorial regression and partial least squares methods. en. *Theoretical and Applied Genetics* **112,** 1009–1023 (Apr. 2006).

57. Velazco, J. G. *et al.* Modelling spatial trends in sorghum breeding field trials using a two-dimensional P-spline mixed model. en. *Theoretical and Applied Genetics* **130,** 1375–1392 (July 2017).

58. Virlet, N., Sabermanesh, K., Sadeghi-Tehran, P. & Hawkesford, M. J. Field Scanalyzer: An automated robotic field phenotyping platform for detailed crop monitoring. en. *Functional Plant Biology* **44,** 143 (2017).

59. Wand, M. P. Smoothing and mixed models. en. *Computational Statistics* **18,** 223–249 (May 2003).

60. Watson, S. Spatial dependence and block designs in spaced plant herbage trials. en. *The Journal of Agricultural Science* **134,** 245–258 (May 2000).

61. Wilkinson, G. N., Eckert, S. R., Hancock, T. W. & Mayo, O. Nearest Neighbour (NN) Analysis of Field Experiments. *Journal of the Royal Statistical Society. Series B (Methodological)* **45,** 151–211 (1983).

62. Williams, E. R. A neighbour model for field experiments. en. *Biometrika* **73,** 279–287 (Aug. 1986).

63. Williams, E. R. & Luckett, D. J. The use of uniformity data in the design and analysis of cotton and barley variety trials. en. *Australian Journal of Agricultural Research* **39,** 339–350 (1988).

64. Williams, K. A. Contemporary Ergonomics 1987: Proceedings of the Ergonomics Society's 1987 Annual Conference, Swansea, Wales, 6–10 April 1987. *American Journal of Occupational Therapy* **42,** 545–545 (1988).

Appendices

Appendix A

Additional informations on computation

A.1 Element-wise product

The element-wise product between two matrix $\mathbf{A}$ and $\mathbf{B}$ is noted $\mathbf{A} \odot \mathbf{B}$ and is defined in the following way:
For two matrices $\mathbf{A}$, $\mathbf{B}$ of same dimensions $n \times m$, the element-wise product is a $n \times m$ matrix where the elements are defined by:

$$(\mathbf{A} \odot \mathbf{B})_{i,j} = (\mathbf{A})_{i,j} \cdot (\mathbf{B})_{i,j}$$

The product is undefined for matrices of different dimensions

A.2 Kronecker product

The Kronecker product of two matrix $\mathbf{A}$ and $\mathbf{B}$ of respective dimensions $n \times m$ and $p \times q$ is a $np \times mq$ block matrix where the elements are defined by:

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & \cdots & a_{1n}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{m1}\mathbf{B} & \cdots & a_{mn}\mathbf{B} \end{bmatrix}$$

A.3 Polynomials splines

Fahrmeir *et al.* (2013) state that a function $f : [a, b] \to \mathbb{R}$ is called a polynomial spline of degree $l \geq 0$ with knots $a = \kappa_1 < \ldots < \kappa_m = b$, if it fulfills the following conditions:

1. $f(z)$ is $(l-1)$ times continuously differentiable. The special case of $l = 1$ corresponds to $f(z)$ being continuous (but not differentiable). We do not state any smoothness requirements for $f(z)$ when $l = 0$.

2. $f(z)$ is a polynomial of degree l on intervals $[\kappa_j, \kappa_{j+1})$ defined by the knots.

Moreover, it can be shown that each polynomial spline of degree l with knots $\kappa_1 < \ldots < \kappa_m$ can be uniquely determined as a linear combination of the $d = l + m - 1$ functions $B_1, \ldots, B_d$, called the *basis functions*, since we can uniquely represent all polynomials splines by using these functions.

A.3.1 B-splines

B-splines are polynomial splines with specific basis functions. B-spline basis functions are constructed from piecewise polynomials that are fused smoothly at the knots to achieve the desired smoothness constraints. More specifically, a B-spline basis function consists of $(l + 1)$ polynomial pieces of degree l , which are joined in an $(l - 1)$ continuously differentiable way. All B-spline basis functions are set up based on a given knot configuration. Using the complete basis, the function $f(z)$ can again be represented through a linear combination of $d = m + l - 1$ basis functions, i.e.,

$$f(z) = \sum_{j=1}^{d} \gamma_j B_j(z).$$

The B-splines of order $l = 0$ can be written as

$$B_j^0(z) = \begin{cases} 1 & \kappa_j \leq z < \kappa_{j+1} \\ 0 & \text{otherwise} \end{cases} \qquad j = 1, \ldots, d-1$$

and the B-splines for higher order l can be written as

$$B_j^l(z) = \frac{z - \kappa_{j-l}}{\kappa_j - \kappa_{j-l}} B_{j-1}^{l-1}(z) + \frac{\kappa_{j+1} - z}{\kappa_{j+1} - \kappa_{j+1-l}} B_j^{l-1}(z).$$

The estimation of a polynomial spline in B-spline representation can be traced back to the estimation of a linear model with a large number of parameters and design matrix

$$\mathbf{Z} = \begin{pmatrix} B_1^l(z_1) & \ldots & B_d^l(z_1) \\ \vdots & & \vdots \\ B_1^l(z_n) & \ldots & B_d^l(z_n) \end{pmatrix}.$$

The linear combination of basis functions can then be written in matrix form

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\gamma}$$

where the coefficient matrix, $\boldsymbol{\gamma}$ can be estimated using least squares.
The estimation of a B-spline fit can be summarized in three steps:

1. We calculate a complete B-spline basis for a given number of knots.
2. The least squares estimate $\hat{\boldsymbol{\gamma}}$ yields an amplitude $\hat{\gamma}_j$ for the scaling of every basis function.
3. We obtain the final estimate by summing the scaled basis function.

A.3.2 Penalized splines

We clearly see that the quality of the estimation by polynomials splines highly depends on the number of knots and that this can easily lead to an over-fitting issue. To overcome this problem, *penalized splines (P-splines)* introduce a roughness penalty term that prevents over-fitting and minimize a *penalized least squares (PLS) criterion* instead of the usual least squares criterion.
To characterize the smoothness of any type of function, the use of (squared) derivatives is appropriate, since these represent measures for the variability of a function. Therefore penalties based on the second derivative, such as

$$\lambda \int \left(f''(z)\right)^2 dz,$$

are particularly attractive since they measure the curvature of a function. Since we know that the first derivative of a B-spline can be written as a function of the first differences of the corresponding coefficient vector, we can use differences of a higher order r if we aim at a smooth function in terms of rth-order derivatives. This leads to the penalized residual sum of squares

$$\mathrm{PLS}(\lambda) = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{d} \gamma_j B_j\left(z_i\right) \right)^2 + \lambda \sum_{j=r+1}^{d} \left(\Delta^r \gamma_j\right)^2,$$

where Δ^r denotes the rth-order differences.The smoothing parameter $\lambda \geq 0$ controls the compromise between fidelity to the data and smoothness of the resulting function estimate. The PLS criterion can be rewritten using matrix notation

$$\mathrm{PLS}(\lambda) = (\mathbf{y} - \mathbf{Z}\boldsymbol{\gamma})'(\mathbf{y} - \mathbf{Z}\boldsymbol{\gamma}) + \lambda \boldsymbol{\gamma}' \mathbf{K_r} \boldsymbol{\gamma}$$

where $\mathbf{K_r}$ is the rth-order difference penalty matrix, and can be decomposed as $\mathbf{D_r}\prime\mathbf{D_r}$ with D_r the rth-order difference matrix. The smoothing parameter $\lambda \geq 0$ controls the compromise between fidelity to the data and smoothness of the resulting function estimate. The PLS estimate of the coefficient matrix is then

$$\hat{\boldsymbol{\gamma}} = \left(\mathbf{Z}'\mathbf{Z} + \lambda \mathbf{K}\right)^{-1} \mathbf{Z}'\mathbf{y}.$$

For more detailed information about polynomials splines, please refer to Fahrmeir *et al.* (2013) and P. H. C. Eilers & Marx (1996)

A.4 Penalized form of the solution

Let us consider the following model, representing the bivariate surface:

$$\hat{\boldsymbol{\alpha}} = (\mathbf{B}^t\mathbf{B})^{-1}\mathbf{B}^t\mathbf{y} \tag{A.4.01}$$

Since the model is purely parametric, it can be estimated by minimizing the residual sum of squares (with explicit solution $\hat{\boldsymbol{\alpha}} = (\mathbf{B}^t\mathbf{B})^{-1}\mathbf{B}^t\mathbf{y}$). To prevent over-fitting, P. H. C. Eilers & Marx (1996) propose to incorporate a discrete penalty on the coefficient associated to adjacent B-splines. For the two-dimensional case, the vector $\boldsymbol{\alpha}$ can be seen

as an $(L \times P)$ matrix of coefficients, $\mathbf{A} = [\alpha_{lp}]$. Now the rows and columns of $\mathbf{A}$ correspond to the regression coefficients in the v and u direction, respectively. In anisotropic (direction-dependant) P-splines, a different amount of smoothing is assumed along the u and v directions. It leads to two penalties: one on all rows of $\mathbf{A}$, the other on all of its columns; and the penalized least squares objective function becomes (P. H. C. Eilers & Marx 2003)

$$\begin{aligned} S^* = & \underbrace{\|\boldsymbol{y} - \boldsymbol{B}\boldsymbol{\alpha}\|^2}_{\text{Original objective function}} \\ & + \underbrace{\hat{\lambda}\|\hat{\boldsymbol{D}}A\|_F^2}_{\text{Penalty along the columns}} \\ & + \underbrace{\breve{\lambda}\left\|\boldsymbol{A}\breve{D}^t\right\|_F^2}_{\text{Penalty along the rows}} \\ = & \|\boldsymbol{y} - \boldsymbol{B}\boldsymbol{\alpha}\|^2 + \boldsymbol{\alpha}^t\boldsymbol{P}\boldsymbol{\alpha}, \end{aligned} \tag{A.4.02}$$

where $\boldsymbol{P} = \hat{\lambda}\left(\boldsymbol{I}_P \otimes \hat{\boldsymbol{D}}^t\hat{\boldsymbol{D}}\right) + \breve{\lambda}\left(\breve{\boldsymbol{D}}^t\breve{\boldsymbol{D}} \otimes \boldsymbol{I}_L\right)$ is the penalty matrix, $\hat{\lambda}$ and $\breve{\lambda}$ are the smoothing parameters acting, respectively, on the columns and rows of $\mathbf{A}$, and $\hat{\mathbf{D}}$ and $\breve{\mathbf{D}}$ are the matrices that form differences of order d_u and d_v respectively. The minimizer of A.4.02 then becomes

$$\widehat{\boldsymbol{\alpha}} = \left(\boldsymbol{B}^t\boldsymbol{B} + \boldsymbol{P}\right)^{-1}\boldsymbol{B}^t\boldsymbol{y}. \tag{A.4.03}$$

Appendix B

Hoagland solution

Table B.1: Composition of the *Hoagland* nutritive solution. The pH must be adjusted to 5.0 using HCl 1% before using.

Components	Concentration (g/L)	ml for 25L of solution[1]
2M KNO_3	202	62.5
2M $Ca(NO_3)_2$ x 4 H_2O	472	62.5
2M $MgSO_4$ x 7 H_2O	493	25
1M NH_4NO_3	80	25
Minors:		
H_3BO_3	2.86	
$MnCl_2$ x 4 H_2O	1.81	
$ZnSO_4$ x 7 H_2O	0.22	25 [2]
$CuSO_4$	0.051	
H_3MoO_4 x H_2O or	0.09	
Na_2MoO_4 x 2 H_2O	0.12	
1M KH_2PO_4 (ph to 6.0 with 3M KOH)	136	12.5
Iron (Sprint 138 iron chelate)	15	75

[1] For a 1:1 solution to use with 25L of water.
[2] All the minors elements are grouped, in the right proportions, in a "minor" solution.

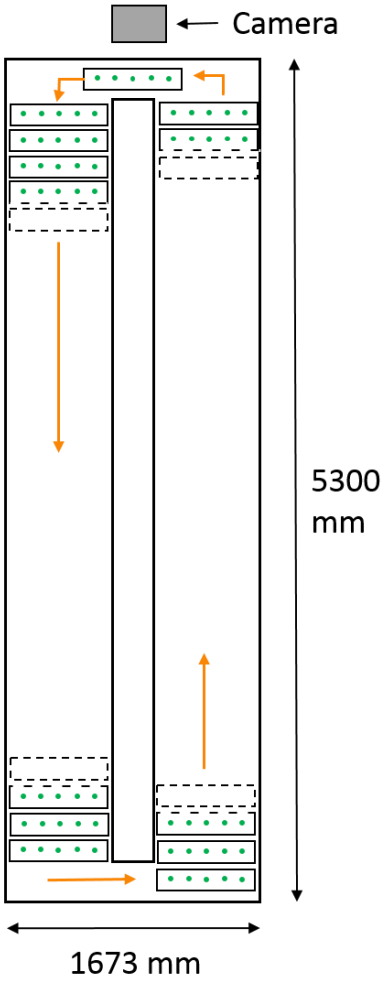
Appendix C

Phenotyping platform information file

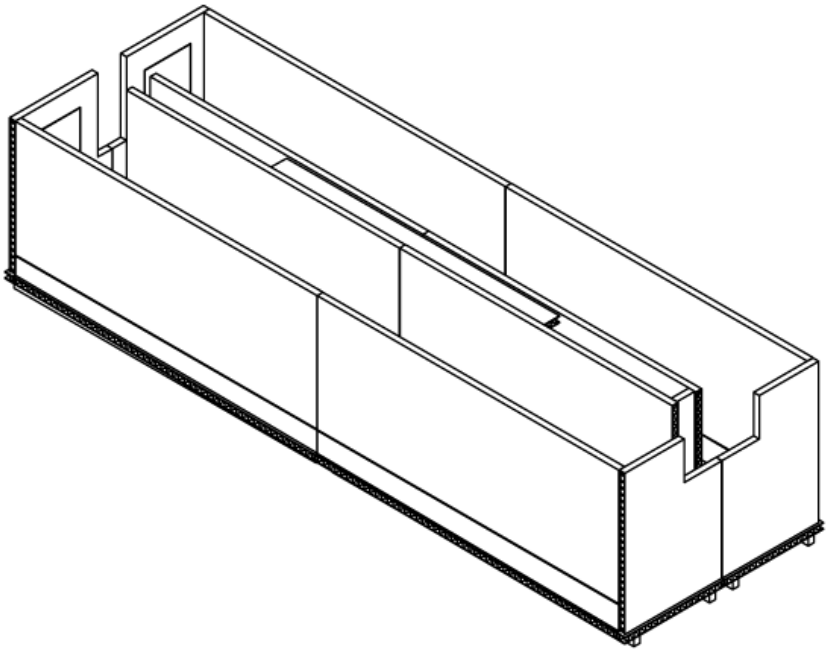
Platform name

Partner site	UCL
Site and installation	Site: Louvain-la-Neuve, Installation: Aeroponics
Contact person(s)	Xavier Draye xavier.draye@uclouvain.be

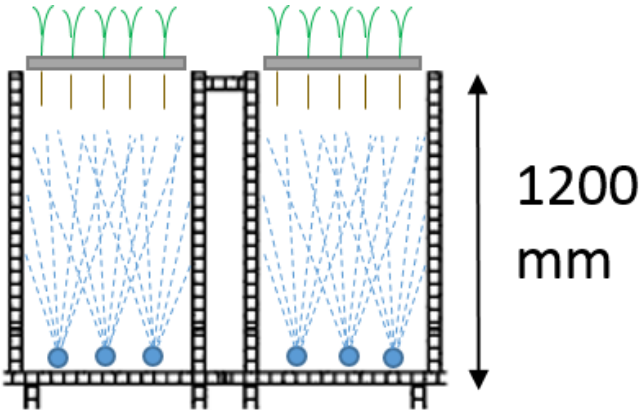
Description of the platform structure



Aeroponic tank: plants are hold on strips, 5 plants per strip (green dots on layout). There are 99 strips in the tank for a total of 495 plants/tank. Strips move in the direction indicated by orange arrows. A full revolution takes 2 hours. When strips pass in front of the camera, at the top of the layout, plants are imaged individually.



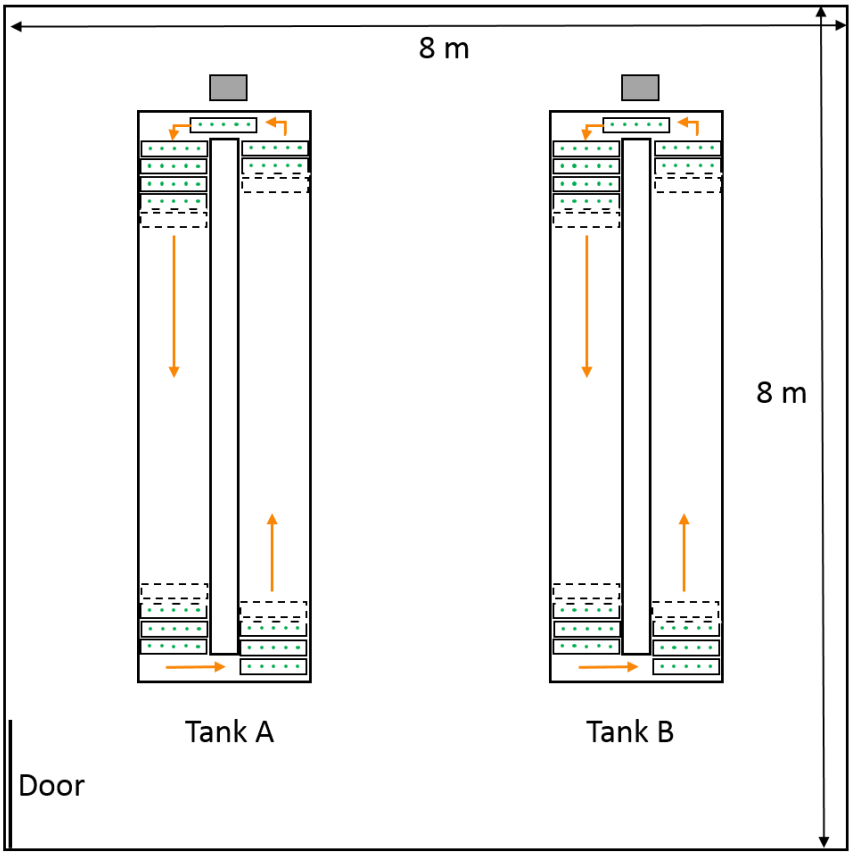
3D view of one tank, without the strips.



Transversal view of the aeroponic tank: 3 sprinklers are placed regularly in the bottom of each side of the tank. The sprinklers spray nutrient solution at regular interval, set by the operator. The spraying

pattern (interval and duration) can be differentiated between day and night and can be modified at any moment of the experiment.

2 identical tanks are available in the installation, located next to each other in the same greenhouse.

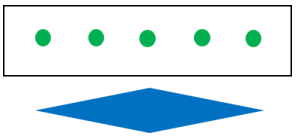


Sources and directions (if known) of environmental variations in the installation

1) Between the 2 tanks.

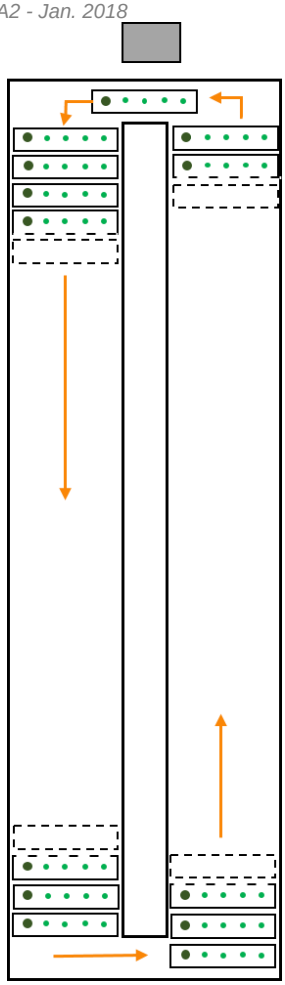
2) The side of the tank placed along the greenhouse wall may be warmer than the side near the centre of the greenhouse because of the presence of heating pipes along the walls.

3) Inside each tank, between plants that grow in the middle of the strip as compared to plants growing at the border of the strip. We suppose that the plants at the extremity of the strip may receive a bit less water than the others.



Layout of a strip with supposed variation of water availability: more water in the middle and less in the border

4) Last year, we observed that the plants growing on the left side of the strips were growing faster than the ones growing on the right side. We understood that the lamps were not exactly centred in the middle of the tank. We moved the lamps to put them exactly at the centre of each tank but we haven't done any new experiment yet.



Layout representing the plants that grow faster on the left side of the strips. The plants keep moving inside the tank but the left/right distinction is maintained during the whole experiment.

As strips keep moving within each tank, we don't expect to observe environmental variation between the different strips of each tank.

Description of experimental design and randomization and a motivation for the design and the randomization

- Design

Completely randomized design: individual plants are located in a strip and at a position randomly with Excel.
+ 2 treatments (eg: shadow, change of nutrient solution properties…) corresponding to the 2 tanks
OR 2 blocks corresponding to the 2 tanks

- Design specifications

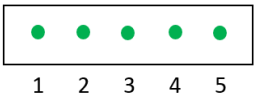
- Motivation

How plant positions are defined and recorded in the experiment

How are the pot positions defined according to the design, i.e. how are the spatial coordinates defined (see example 6)?

QR code associated to each plant

Number of the QR code:
Ex: B_76_5
B: tank id (A or B)
76: strip id (from 1 to 99)
5: position in the strip (from 1 to 5)



If pots are rearranged during the experiment, how is the change in spatial position recorded?

All strips move at the same pace. Each plant passes every 2 hour in front of the camera, where a picture is taken. The time of the picture enables to record the moment at which each plant passes in front of the camera. It would be possible to compute the pathway the plant had in the tank between two pictures.
No changes between the two tanks or within each strip (position 1 to 5)

If repeated measurements are taken, at what times are these taken?

Every 2 hours, 24h a day

Leuven Statistics Research Centre (LStat)
Celestijnenlaan 200 B
3001 HEVERLEE, BELGIË
tel. +32 16 32 88 75
https://lstat.kuleuven.be/contact