

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

© Copyright by KU Leuven

Without written permission of the promotor and the authors it is forbidden to reproduce or adapt in any form or by any means any part of this publication. Requests for obtaining the right to reproduce or utilize parts of this publication should be addressed to KU Leuven, Faculteit Wetenschappen, Geel Huis, Kasteelpark Arenberg 11 bus 2100, 3001 Leuven (Heverlee), Telephone +32 16 32 14 01.

A written permission of the promotor is also required to use the methods, products, schematics and programs described in this work for industrial or commercial use, and for submitting this publication in scientific contests.

Chapter 1

Literature review

1.1 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 clearly¹. We can define the phenotype as the set of all types of traits of an organism or one of its subsystems. It is often confused with the phenotype which is the physical totality of all traits of an organism or one of its subsystems. The same distinctions can be made between the genome and the genotype, where the genome is the physical totality of all genes of a cell, tissue or organism and the genotype is only a set of those genes, distinguished by specific base sequence, or locus (**mahner1997exactly**). We can then re-express the phenotype of an organism as the expression of the genotype as traits in a given environment ($\text{Phenotype} = \text{Gene} \times \text{Environment}$), since the genotype, the environment, and their interaction influence quantitative traits in a complex and dynamic manner.

Therefore, 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 (**houle_phenomics: 2010; fiorani_future_2013**). 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_future_2013**).

Phenotyping is more largely part of phenomics, which is defined as the study of plant growth, performance and composition. Forward phenomics uses phenotyping tools to sieve through the available "genetic pool" of a species (the germplasm) for valuable traits. Reverse phenomics is the detailed dissection of traits to reveal mechanistic understanding and often involves reducing the trait to a biophysical processes and ultimately to genes (**furbank_phenomics_2011**).





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 (**pieruschka2019plant**). The demographic projections show that cereal grain yields must increase by at least 70%

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

(**furbank2009c4**) or even double (**tilman2011global**) before 2050 to meet the predicted production demands of the global population. Even though extensive breeding has been responsible for tripling cereal yield in the last 50 years (**pingali2012green**) and production rates are still increasing yearly, annual increases in yield achieved from traditional breeding programs worldwide are no longer sufficient to meet projected demand for all three major cereal crops: rice (*Oryza sativa*), maize (*Zea mays*) and wheat (*Triticum aestivum*) (**tester2010breeding**). In the actual context of climate change, demand for biofuel feedstocks will also undoubtedly increase over the next decade (**sticklen2007feedstock**), resulting in potential competition for arable land between food and fuel crops. At the same time the impacts of climate change on global temperatures and rainfall patterns are likely to lead to reductions in yields due to abiotic stress (**tester2010breeding**). Therefore, the genetic improvement of crops is imperative to achieve high intrinsic yields and yield stability under abiotic stress, to further ensure future food security. Continuing advances in the techniques available to breeders offer the potential to increase the rate of genetic improvement (**phillips2010mobilizing**), and advances in phenotyping are the key to create and offer new methods that will solve the problems that agriculture is currently facing.

However, plant phenomics does not consist of solely associating a genotype to one phenotype in a given condition (e.g., in a controlled environment), but rather in characterizing the plasticity of the plant phenome when exposed to a range of environmental conditions (**tardieu_plant_2017**). Therefore phenotyping is often a complex task because the plasticity of phenotypes in plants is enormous and plant phenotypic responses are generally characterized by response curves or reactions to the environment, which for complex traits are inherently continuous and mostly non-linear (**sultan2003phenotypic**). Indeed, plants are not homeostatic for temperature and water under rapidly changing conditions. For example, plant evapo-transpiration can range from 50% to 200% of their own weight (**vadez2014transpiration**), leaf temperature can vary by more than 20° in a single day, leading to spectacular changes in plant morphology (**caldeira2014hydraulic**). As a consequence, phenotyping also encompasses the characterization of the environment of the plant, but it has long been difficult to record both biological systems and their environment in the multitude of spatial and temporal dimensions that are necessary to understand the development of a specific phenotype (**pieruschka2019plant**). Mainly because environmental factors influencing plant growth and development are characterized by different levels of heterogeneity in space and time (**hodge2004plastic**). Another source of complexity is the fact that plants do not have a central organ that acts as a control unit for the entire organism. The control of functions often relies on feedback loops involving exchanges of information. Such exchanges of information operate at short-term scales at the cell or organ levels, and translate into long-term plant or canopy behaviours through whole-plant mechanisms that are highly non-linear. Hence, plant phenomics requires analyses at spatial scales ranging from single cell to canopy, and temporal scales ranging from minutes (for metabolism and hydraulics) to months (for yield) (**tardieu_plant_2017**). Table 1.1 gives a quick overview of the different scales present in phenotyping and their associated methods and mechanisms.

Table 1.1: Phenotyping at different scales of organization. Redrawn from **tardieu_plant_2017**

Typical phenotyping platforms	High-precision platforms (omic, anatomy, organ)	Whole-plant, multi-environment platforms (field or controlled)	Field multi-environment networks
Level of plant organization	Organ	Plant or canopy	Canopies in a range of environments
Typical methods	Omics, 4D organ imaging, fluxes	4D plant/canopy imaging fluxes, sensors	Yield, sensor network, remote sensing
Typical mechanism	Hydraulics, metabolism, signalling	Light interception, water transfer, whole-plant signalling	Trade-offs between processes
Ratio biology/other processes			
Relevance for yield prediction			
Methods for trans-scale communication	<div style="display: flex; justify-content: space-around; align-items: center;"> <div style="text-align: center;">  <p>Gene editing, plant simulation</p> </div> <div style="text-align: center;">  <p>GWAS, model-assisted dissection</p> </div> </div>		

1.1.1 Phenotyping bottleneck

Currently, the rate at which phenotypes are extracted in the field or in the lab is not matching the speed of genotyping (**houle_phenomics: 2010**). While genetic editing techniques and genome mapping technologies are blooming, they depend on a similar improvement in phenotyping, since they are key to analyze plant responses to environmental characterization. This issue is creating a phenotyping bottleneck.

In the recent years, there has been a significant advance in the large-scale characterization of plant genomes. This is mainly due to advances in molecular profiling techniques such as marker-assisted selection (MAS) and genomic selection tools for specific alleles (**tester2010breeding; jannink2010genomic**), but also to genome sequencing technologies that allow fast and cheap DNA sequencing (**yano2009genome**). The use of model plants, such as *Arabidopsis thaliana* (**atwell2010genome**), and the introduction of suitable statistical packages and bioinformatics tools has lead to fast and inexpensive genomic information. This fast-paced, high-throughput and low-cost genotyping has paved the way for the development of large mapping populations ² (**mcmullen2009genetic**).

However, similar advances in phenotyping are now needed to capitalize on those developments and ensure genetic improvement of crops for future food security (**araus_field_2014**).

As said previously, the plasticity of the phenotype according to its environment and the heterogeneity of said environment, make it complex to capture plant phenotype in a dynamic and significant way. While this barrier was previously concerning the platforms and equipment (the hardware), non-invasive methods for plant characterization are now blooming (**li_review_2014**). The new weak spots in the phenotyping chains are now the analysis of imaging data (the software), and the need for a comprehensive approach and good data management (**fiorani_future_2013**).

Image analysis is an issue due to the challenges in computer visions and data processing. Plants are self-changing systems with increasing complexity over times and spatial scales that vary from the subcellular level to the large outdoor field. Therefore, image capturing processes go from automatic cell delineation at the microscopic level, to complete 3-D shape reconstruction at the whole plant level (**minervini2015image**). Moreover, the rapid progress that has been made in the development of a wide array of technologies including novel sensors, image analysis, data mining and modelling, now allow plot-level measurements within seconds. The largest limitation to the implementation of phenotyping platforms is not time anymore, but the management of the huge amount of information generated (**cobb2013next; aaraus_field_2014**). Image capturing and processing algorithms must deal with this quantity of complex and diverse raw information, and must also be efficient enough to allow automation and create high-throughput phenotyping platforms (HTPP).

fiorani_future_2013 also specify that the new innovative technologies need to be integrated in a broader multi-faceted approach, to help relieve the phenotyping bottleneck. The same authors consider a generic process scheme for phenotyping in indoor growth facilities in figure 1.1. They subdivide the phenotyping process in 3 layers: the

²A population that is suitable for linkage mapping of genetic markers is known as mapping population. Mapping populations are generated by crossing two or more genetically diverse lines and handling the progeny in a definite fashion. Generally, the parents used for hybridization will be from the same species (**singh2015mapping**).

experimental design layer, is about having adequate, high-capacity, infrastructures for phenotyping to allow high-throughput; the automated analyses layer, encapsulates the need for standardized and well-defined protocols for data collection and analysis, to allow reproducibility; the last layer emphasizes the need for good software for numerical and statistical analysis and metadata for results interpretation. This scheme also describes the need a good understanding of the parameters driving the evolution of the plant characteristics researchers are measuring. It is clearly valuable to identify a set of parameters before wasting resources by measuring a large number of data points, which could be highly autocorrelated or not indicative of the target performance.

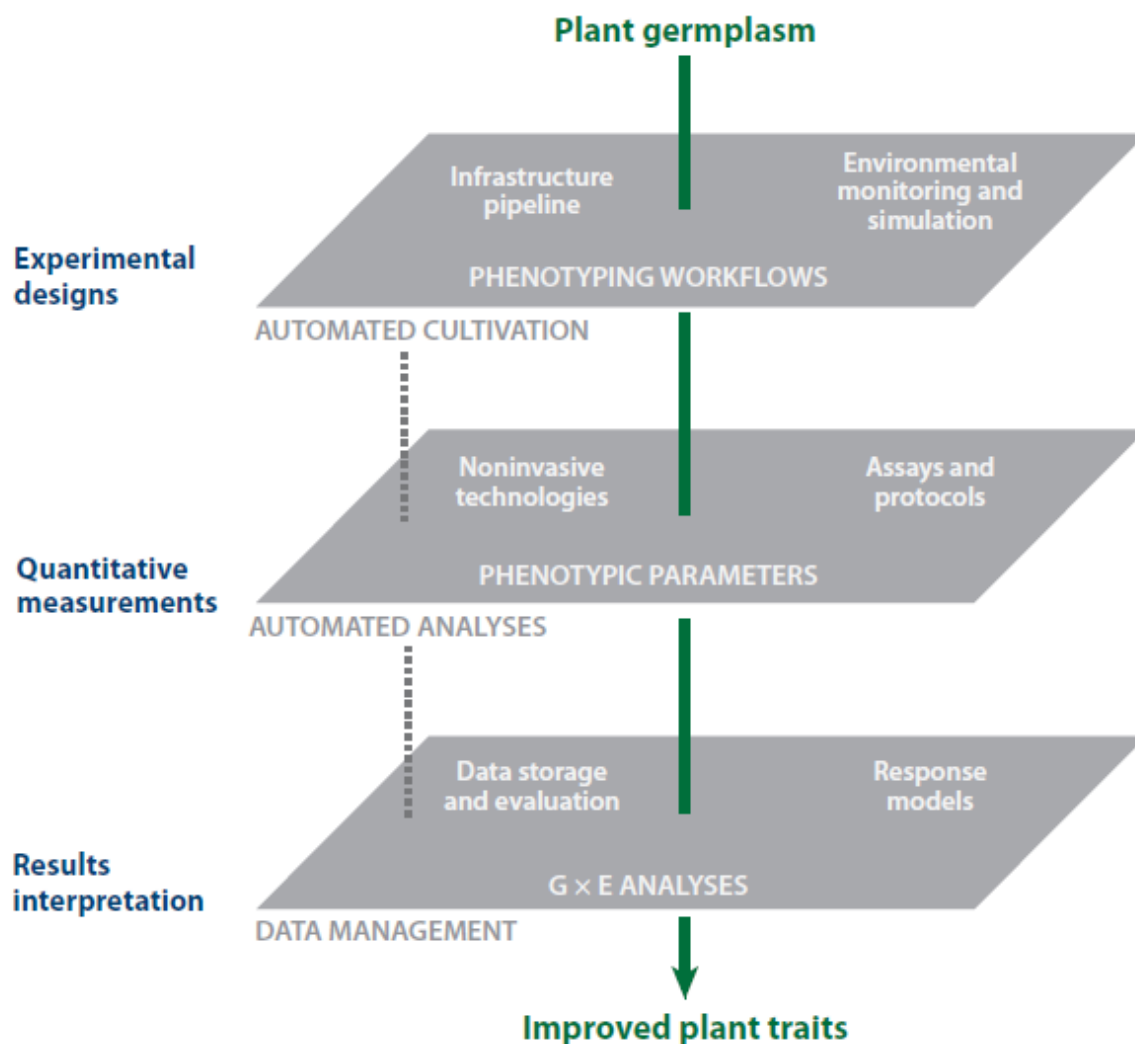


Figure 1.1: Conceptual scheme for plant phenotyping, applicable in particular to controlled-environment facilities. Building capacities to screen germplasm for enhanced agricultural traits requires a multidisciplinary approach. Globally, this scheme offers a quick overview of key layers and elements for a successful implementation of large-scale plant phenotyping. This architecture implies a direct link between the measured plant parameters and the environmental conditions, enabling the analysis of gene-environment ($G \times E$) interactions and modeling of phenotypic responses (fiorani_future_2013).

1.1.2 Data processing and standardization in phenotyping

1.1.3 Community integration and collaboration

1.2 Experimental design in field trials

Experimental field trials in agriculture have always been affected by soil heterogeneity. As **van_es_1.2_2002** explains, soil is a continuum with variability on multiple scales. The heterogeneity is as much affected by microscopic interactions than 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_one_2001**). For randomisation to be truly effective, stationarity of the mean and spatial independence assumptions need to be verified. Several studies have proved that it is rare that both these assumptions hold in field trials (**davidoff_method_1986**; **nielsen_spatial_1973**; **iqbal_spatial_2005**). Moreover, **van_es_spatial_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_design_2006**; **velazco_modelling_2017**). Local field trends can influence groups of treatments in specific blocks. As a solution, several authors (**watson_spatial_2000**; **fagroud_accounting_2002**) 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_spatially-balanced_2007** showed that completely randomized (43 % in greenhouse trials) and random block designs (70 % in field trials) are still vastly used. Considering this global issue, finding and using an appropriate design is complex task.

1.3 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_spatial_1991**) have become a standard strategy in field trials. However, **piepho_problems_2015** recently discussed several issues with this model and have therefore proposed the use of the linear variance (LV) model (**williams_use_1988**) instead. More specifically, **piepho_linear_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, **rodriguez-alvarez_correcting_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 the penalized splines (**eilers_flexible_1996**) to esti-

mate a bivariate smooth function over the rows and columns of a plot. Using the work of **lee_efficient_2013**; **lee_hwang_smoothing_2010**; **lee_p-spline_2011** the spatial variability is characterized using tensor products of two-dimensional P-splines (**dierckx_curve_1995**) and decomposed in a PS-ANOVA system. By exploiting the similarities between P-splines and mixed models (**currie_flexible_2002**; **durban_adjusting_2001**; **wand_smoothing_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 (**rodriguez-alvarez_correcting_2018**) and previous field trials data (**lado_increased_2013**) and showed promising results.

As **wilkinson1983nearest** highlight, in field trials data modelling, the main components of spatial variation are:

- non-stationary large scale (global) variations across the field ³
- stationary variation within the trial (natural variation or local trend)
- extraneous variations (often due to experimental procedures).

The spatial variation can also be attributed to systematic effects, e.g. sowing or planting, and 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. Since the spatial variation has both random and systematic components, it is straightforward use the mixed model framework.

Random spatial variations are harder to model because there are no covariates to relate it to. There are two main approaches to model spatial trends: one based on spatial variance-covariance structures; and the other based on smoothing techniques. Here the yield data, extracted from the phenotyping platform, are modelled using these 2 different models.

1.4 Thesis objectives

This master thesis falls within the scope of the second activity of the European project EPPN2020⁴. 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

³**risser2016nonstationary** 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 $\mathbf{h} \in \mathcal{R}^d$.

⁴European Plant Phenotyping Network 2020 <https://eppn2020.plant-phenotyping.eu/>

revolves around data acquisition, data analysis and data networking, so that every platform use common, standardized practices and analysis protocols, that have been verified for robustness and quality.

The main goal is 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 fitted using JMP, taking into account the fact that few replications of the trial will be made and the number of genotype tested. Its efficiency will be tested against classical pre-made designs, such as the Randomized Complete Block (RCB) and alpha-lattice designs, and the best one will be applied on the phenotyping platform. 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 **piepho_linear_2010**. The second one is the SpATS (Spatial Analysis using Tensor product of Splines) model, recently created by **rodriguez-alvarez_correcting_2018**. These models can be compared using classical precision measures (RMSE,...) but also through more relevant ones like effective dimensions and heritability (**oakey_joint_2006**).

The experiment will take place in January in the UCLouvain greenhouses ⁵. The installation consists of two aeroponic tanks of 495 plants located in a 64 m² G2 greenhouse. The greenhouse is equipped with temperature and humidity sensors to monitor the local atmosphere. Plants are held on strips, 5 plants per strip, 99 strips per tank. Sprinklers placed at the bottom of the tanks spray nutrient solution on the roots of the 495 plants. Sprinkling solutions and patterns can be customized and even modified during the experiment if needed. Strips move constantly and plants are scanned when the strip passes in front of a camera. This camera consists of a high resolution scanner that captures a full 2D projection of the roots and shoots of each plant. As a full tank revolution takes two hours, plants are imaged every two hours. To ensure minimal manipulation, seed germination occurs within the platform. Experiments usually last 3 weeks, which is the time it takes to the root system to reach a depth of 60 cm. The experiment will include two tanks. In the first one, plants will constantly move to be pictured every 2 hours (usual setup on this platform). In the second one, plants will move twice or three times a day to be pictured. This will allow comparing the effect of moving vs. 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 will be the overall growth of the root system of each plant. An experiment generates a large amount of data in a raw, image-based format (ca. 200K images). Images allow temporal decoupling, provide a condensed set of information and are multi-dimensional (2D usually, but 3D scanning platforms are common (**mooney_developing_2012**)). A lot of tools are available for data analy-

⁵A detailed plan of the installation is attached to this document.

sis in phenotyping platforms (**lobet_online_2013**). This makes the choice complicated for an external user, especially since most of these software 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 (**pound_rootnav: 2013; lobet_novel_2013**). Scientists of the UCLouvain platform have developed pipelines⁶ that allow easy processing of the images captured in the platform to extract quantitative root architecture information for the spatial models (**lobet_novel_2013; lobet_novel_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.

⁶Here, pipelines are defined as computer programs designed to analyse raw data from phenotyping platforms.

Chapter 2

Material and methods

2.1 Optimal experimental designs

Say how custom designs are more efficient because they fit to the problem rather than having to 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.

2.1.1 Orthogonal designs

Consider a regression model written in matrix form

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad (2.1.11)$$

where \mathbf{Y} is a $(nx1)$ column vector for the response variable, \mathbf{X} is a $(n \times p)$ design matrix, $\boldsymbol{\beta}$ is a $(p \times 1)$ column vector for the coefficients and $\boldsymbol{\epsilon}$ is a $(nx1)$ column vector for the random error term.

The best linear unbiased estimator (BLUE) of the coefficients matrix is given by the ordinary least square (OLS) estimator

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}, \quad (2.1.12)$$

because it minimizes the variance of the estimators

$$\text{var}(\hat{\boldsymbol{\beta}}) = \sigma_{\epsilon}^2(\mathbf{X}'\mathbf{X})^{-1}, \quad (2.1.13)$$

where $\sigma_{\epsilon}^2 = \text{Var}(\epsilon_i)$ is the variance of the random error term ϵ . In its developed form, the variance-covariance matrix is

$$\text{var}(\hat{\boldsymbol{\beta}}) = \begin{bmatrix} \text{var}(\hat{\beta}_0) & \text{cov}(\hat{\beta}_0, \hat{\beta}_1) & \vdots & \text{cov}(\hat{\beta}_0, \hat{\beta}_k) \\ \text{cov}(\hat{\beta}_1, \hat{\beta}_0) & \text{var}(\hat{\beta}_1) & \vdots & \text{cov}(\hat{\beta}_1, \hat{\beta}_k) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(\hat{\beta}_k, \hat{\beta}_0) & \text{cov}(\hat{\beta}_k, \hat{\beta}_1) & \dots & \text{var}(\hat{\beta}_k) \end{bmatrix}, \quad (2.1.14)$$

where the diagonal elements represent the individual variances of the estimates of the model coefficients and the off-diagonal elements represent the covariances between

pairs of estimates. Ideally, these variances should be as small as possible to allow powerful significance tests and narrow confidence intervals about the unknown factor effects.

When planning an experiment, the error variance σ_ϵ^2 is usually unknown and often set to one to only consider the elements of $(\mathbf{X}'\mathbf{X})^{-1}$. The diagonal elements of that matrix are called the relative variances of the estimates of the model's parameters and are denoted by

$$v_i = \frac{1}{\sigma_\epsilon^2} \text{var}(\hat{\beta}_i). \quad (2.1.15)$$

The inverse of the variance-covariance matrix (2.1.14), denoted by

$$\mathbf{M} = \frac{1}{\sigma_\epsilon^2} \mathbf{X}'\mathbf{X}, \quad (2.1.16)$$

is called the information matrix because it summarizes the available information on the model's parameters.

As detailed by **goos2011optimal**, when the $(\mathbf{X}'\mathbf{X})$ matrix is diagonal and when all the diagonal elements are n then the design is said to be orthogonal.

An orthogonal design allows for the independent estimation of the model's parameters and the covariance between each pair of estimates is zero. Therefore the estimates are said to be independent and uncorrelated. Moreover when a design is orthogonal, the addition or subtraction of terms in the model does not affect the estimates.

Finally, the information matrix (2.1.16) is also diagonal whenever the design is orthogonal.

Variance inflation

When a design is not orthogonal, the relative variances (2.1.15) of at least one of estimates of the model parameters will be bigger than $1/n$. The variance is said to be inflated and the factor by which it is inflated is called the variance inflation factor (VIF). For two-level designs, the VIF is

$$\text{VIF} = nv_i \quad (2.1.17)$$

The minimum value for the VIF is one. For orthogonal designs, the VIF of every factor-effect estimate is one. The rule of thumb used by many authors, is that a VIF of more than 5 indicates potential collinearity problems.

2.1.2 Optimality criteria

In order to generate optimal designs, one needs to use an optimality criterion to compare different designs together. 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 the G-optimality criterion that is similar to the I-optimality criterion but minimizes the maximum prediction variance. Recent work (**rodriguez2010generating**) has shown that I-optimal designs are often better choices than the G-optimal ones.

D-optimality

As said previously, for an orthogonal design all the non-diagonal elements of the $(\mathbf{X}'\mathbf{X})^{-1}$ matrix are null and the determinant of the matrix, $|\mathbf{X}'\mathbf{X}|^{-1}$ is simply the product of the diagonal elements. Since the goal is to have the smallest value for the variance of the estimates, an orthogonal design will have the smallest determinant value. Therefore the determinant can be used as an overall measure of the variance of the estimates.

Minimizing the determinant of the $(\mathbf{X}'\mathbf{X})^{-1}$ is similar to maximizing the determinant of the $(\mathbf{X}'\mathbf{X})$ matrix. Since this matrix is proportional to the information matrix (2.1.16), the factor settings that maximize the determinant of $(\mathbf{X}'\mathbf{X})$ will maximize the available information about the model's parameters. The design that maximize the determinant of the $(\mathbf{X}'\mathbf{X})$ matrix 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. As mentioned in section 2.1.1, in non-orthogonal designs, the variance of the estimates is inflated and they are correlated. However this inflation is usually small and not dramatic 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 $(\mathbf{X}'\mathbf{X})$ matrix.

D-efficiency The D-efficiency of a given design compares the determinant of the information matrix of that design to an ideal determinant corresponding to an orthogonal design. Since orthogonal design do not always exist, the ideal determinant is defined as n^p for a design with n runs and p parameters in the vector β . The D-efficiency is therefore computed as

$$\text{D-efficiency} = \left(\frac{|\mathbf{X}'\mathbf{X}|}{n^p} \right)^{1/p} = \frac{|\mathbf{X}'\mathbf{X}|^{1/p}}{n}, \quad (2.1.21)$$

where the p th root is taken to provide a measure that can be interpreted as a per-parameter measure. The two main problems of the D-efficiency is that it depends on the coding scale of the parameters and that the ideal determinant cannot always be achieved, therefore biasing the measure. The D-efficiency is then only useful to compare design that have the coding for the experimental factors and the same number of runs. For this reason, it is more useful to compare the relative efficiency of two designs it's better to use the relative D-efficiency, defined as

$$\text{Relative D-efficiency of Design 1 vs Design 2} = \left(\frac{D_1}{D_2} \right)^{1/p}. \quad (2.1.22)$$

I-optimality

Variance of prediction The variance of the prediction's expectation at the setting \mathbf{x} (a vector of factor levels that corresponds to one of the runs of the design) is

$$\text{var}(\hat{Y}|\mathbf{x}) = \sigma_\epsilon^2 \mathbf{f}'(\mathbf{x})(\mathbf{X}'\mathbf{X})^{-1}\mathbf{f}(\mathbf{x}), \quad (2.1.23)$$

where $\mathbf{f}(\mathbf{x})$ is a function that takes a vector of factor settings and expands it to its corresponding model terms. The variance expressed relatively to the error variance σ_ϵ^2 , is

$$\text{Relative variance of prediction} = \frac{\text{var}(\hat{Y}|\mathbf{x})}{\sigma_\epsilon^2} = \mathbf{f}'(\mathbf{x})(\mathbf{X}'\mathbf{X})^{-1}\mathbf{f}(\mathbf{x}). \quad (2.1.24)$$

Since the prediction's variance depends on the factor settings, \mathbf{x} , it is interesting to calculate the averaged variance of prediction over the design space. The average is computed by integrating the relative variance of prediction (2.1.24) over the experimental region χ , and dividing it by the volume of the region:

$$\text{Average variance} = \frac{\int_\chi \mathbf{f}'(\mathbf{x})(\mathbf{X}'\mathbf{X})^{-1}\mathbf{f}(\mathbf{x})d\mathbf{x}}{\int_\chi d\mathbf{x}}. \quad (2.1.25)$$

A design that minimizes the average relative variance of prediction is called a "I-optimal" design. The "I" in "I-optimal" stresses out the fact that an I-optimal design minimizes an integrated variance¹.

I-efficiency Similar to the relative D-efficiency, the relative I-efficiency allows for the comparison of two designs in term of average relative variance of prediction. If P_1 and P_2 are the average relative variance of prediction for two different design, then the relative I-efficiency of the first design to the other is computed as

$$\text{Relative I-efficiency of Design 1 vs Design 2} = \frac{P_1}{P_2}. \quad (2.1.26)$$

2.1.3 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 **meyer1995coordinate**. 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 **fedorov2013theory** and modified several times to speed it up (**johnson1983some**; **atkinson1989construction**). 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 (**heredia2003genetic**; **heredia2004model**), simulating annealing algorithms (**bohachevsky1986generalized**; **meyer1988constructing**) and tabu search

¹Some authors prefer the term IV-optimal or V-optimal over I-optimal.

algorithms (**jung1996construction**) 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 rank of the matrix of factors settings

$$\mathbf{D} = \begin{bmatrix} x_{11} & x_{21} & \dots & x_{k1} \\ x_{12} & x_{22} & \dots & x_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1n} & x_{2n} & \dots & x_{kn} \end{bmatrix}, \quad (2.1.31)$$

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 the set $\{-1, +1\}$. 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.

2.1.4 Generating the design

A custom design was fitted to the experiment and four factors were considered:

Tank In which tank was situated the plant (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).

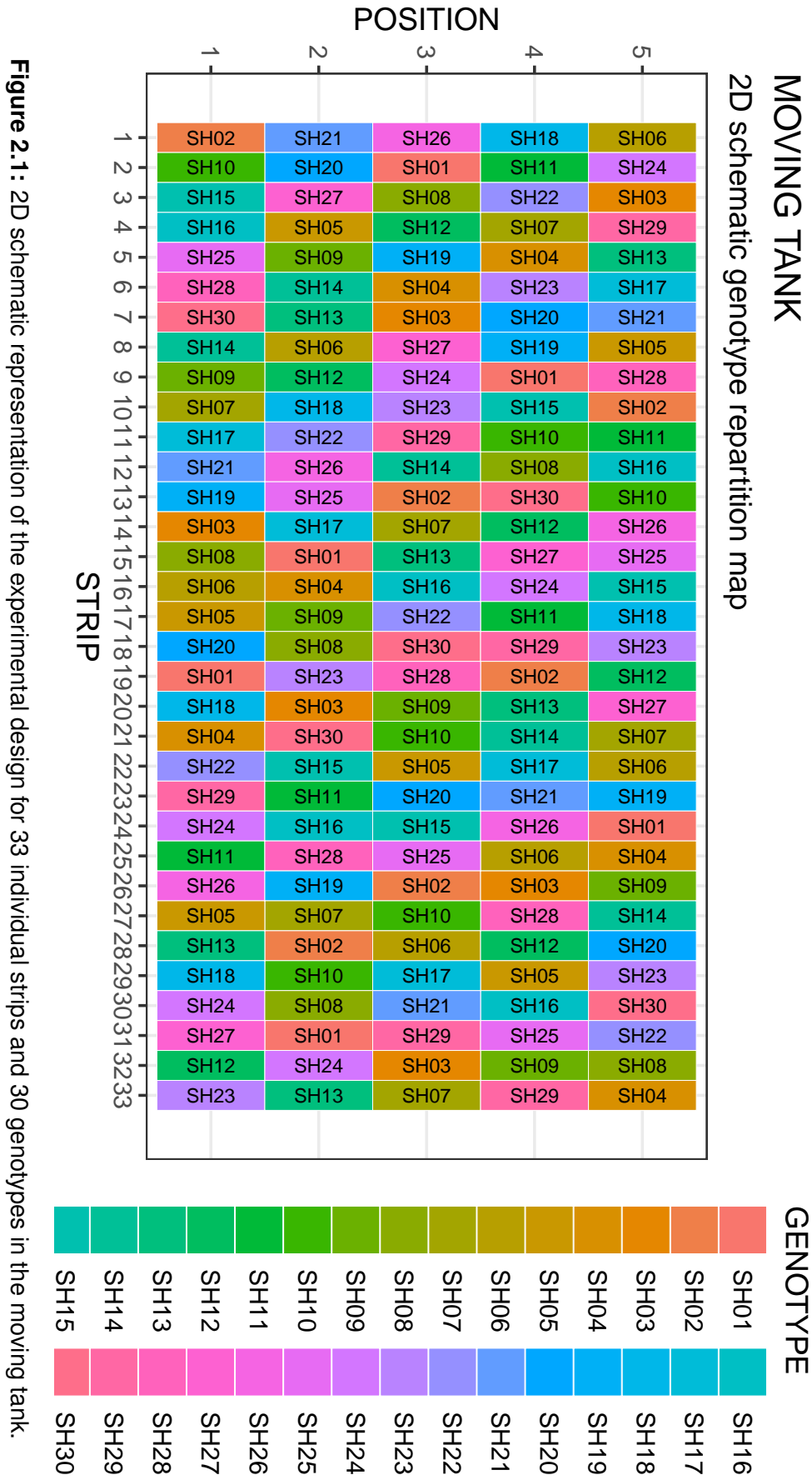
Genotype Which one of the 30 genotypes was used (1 to 30).

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, there were 990 experimental runs 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 2.1 and 2.2, 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 series². 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) will be occupied, the border genotype will also be used to fill the 90 empty runs.

²Historical series correspond to varieties that have been cultivated and bred for some time, mainly due to their physiological specificities.



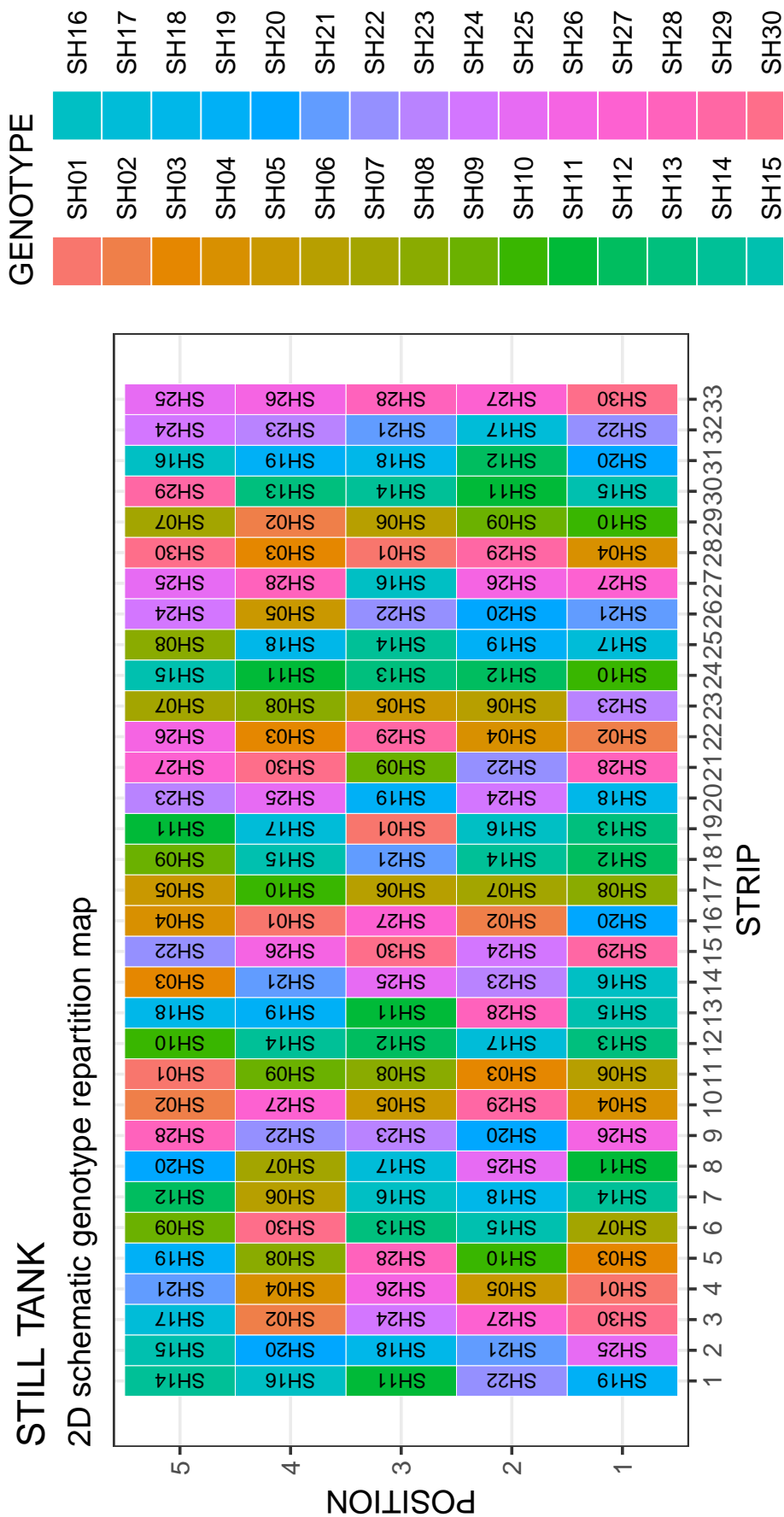


Figure 2.2: 2D schematic representation of the experimental design for 33 individual strips and 30 genotypes in the still tank.

2.2 Phenotyping experiment

The phenotyping experiment took place during four weeks between February and March. 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.

2.2.1 Germination

A previous germination experiment was conducted in the greenhouse, in a smaller version of the phenotyping platform, to compare seed germination in different type of substrate. In this experiment, 60 maize seeds were set on 6 strips of 5 positions in 2 small reproductions of the aeroponic tanks of the platform. In each tank, a different kind of cork was used to see which one allowed a better germination. It showed that germination rates were fairly low (around 6%) for both type of cork, mainly because the seeds were asphyxiated in the corks. This lead to the choice of an outside germination to avoid low germination and seed asphyxiation problems.

The seeds were germinated in a germination chamber outside of the platform and were transferred onto the platform once they were germinated.

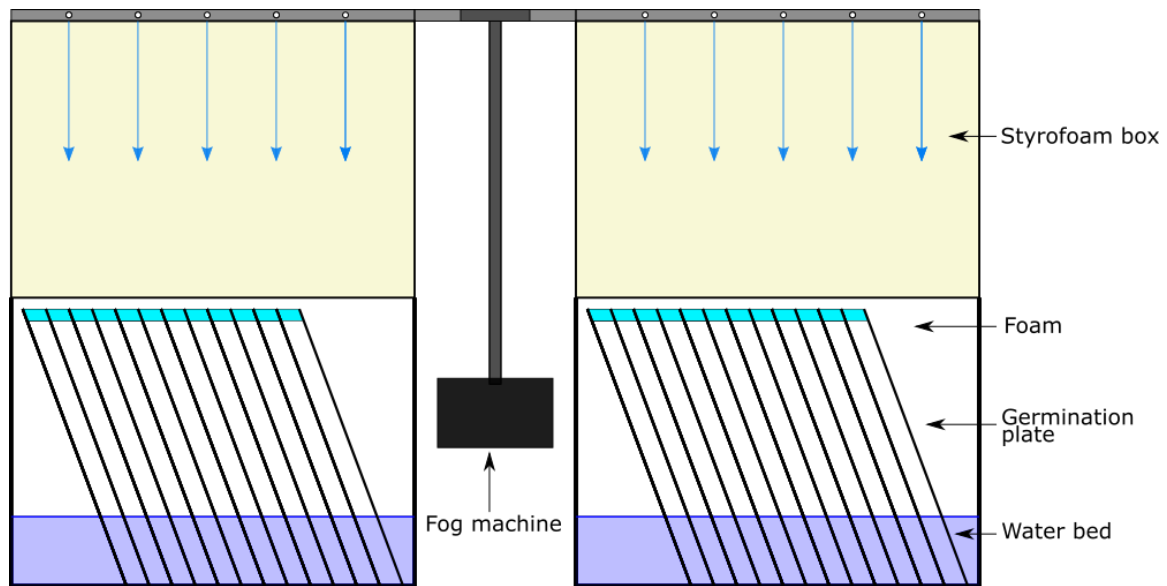
Germination chamber

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 was connected an air-fog machine to keep the seeds moist. Inside each tray, fifteen 40cm by 30cm PVC plates were disposed diagonally and evenly spaced (fig. 2.3a). On each PVC plate, the seeds were arranged on a filtering paper sheet with ledges to support the weight of the seeds (figure 2.3b). All the plates were set to a 60° angle and 5 cm apart in the PVC tray with 5 cm of water in the bottom 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). Each sheet had 6 rows of 10 seeds with one genotype on the left and one genotype on the right, see figure 2.3c for details. The seeds were attached with an additional drop of agar solution (1% in volume) to stabilize their position and attach them to the paper.

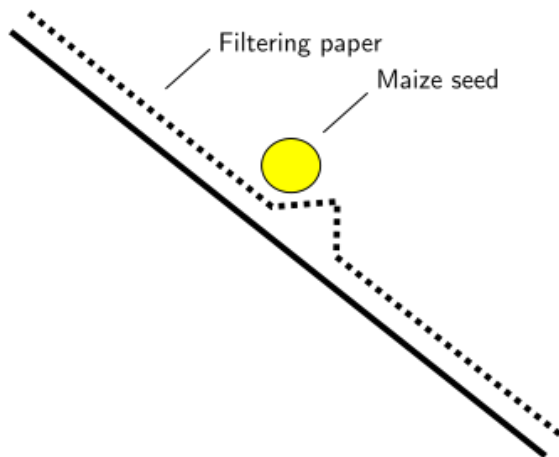
After 3 days into the chamber, not all seeds were germinated, table 2.1 presents the germination rates and mean seed weights for all the genotypes used (including the border genotype). The non-germination was mainly due to the fact that seeds fell into the water bed (due to a lack of support on the sheet because there was not enough agar solution). Some other seeds did not germinated because mold grew on the filtering paper and on the agar drops, even though the drops were set in a sterile environment.

2.2.2 Phenotyping platform

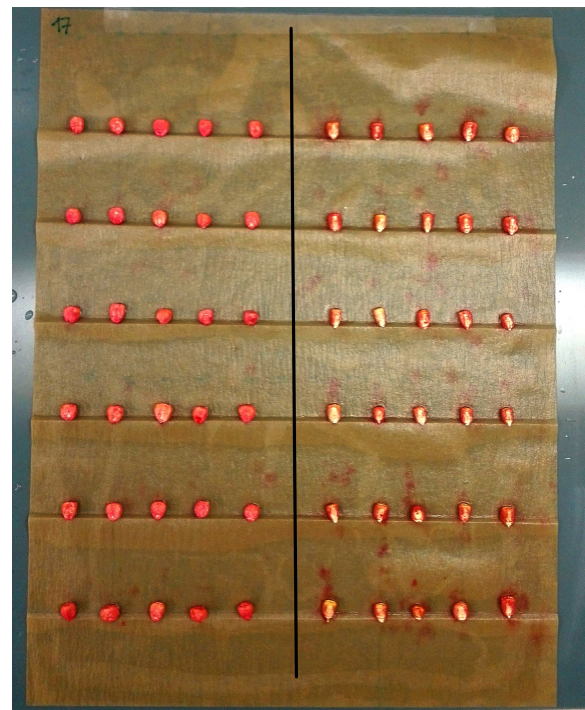
The phenotyping platform is located inside a 8m by 8m greenhouse in the facilities of the UCLouvain (Louvain-la-Neuve, Belgium). It consists of two aeroponic tanks of 1673



(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)



(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

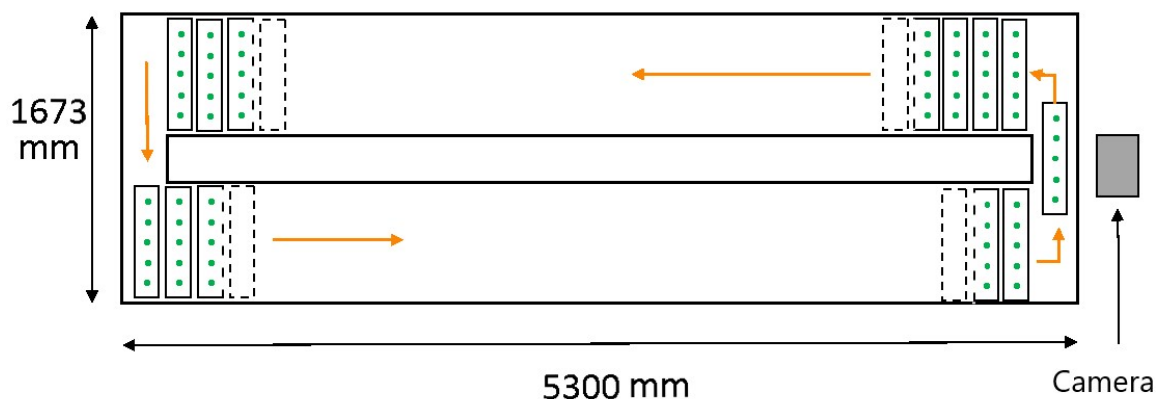


(c) PVC plate with seeds on filtering paper (the black line represents the separation between the two genotypes on the plate)

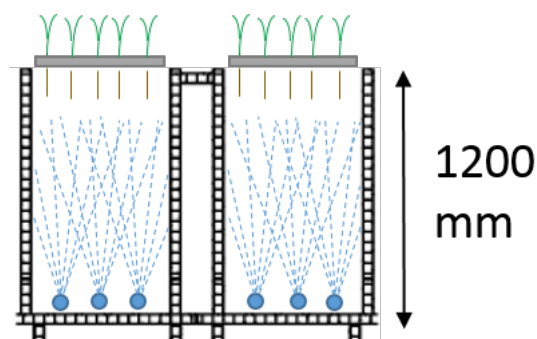
Figure 2.3: Germination chamber diagram with detailed view and pictures

mm by 5300 mm on which are arranged 99 styrofoam strips, each with five holes (\varnothing 2.5 cm). 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. 2.4a). The strips rotate clockwise in the tank and a full rotation is completed in 2 hours. Three sprinklers are

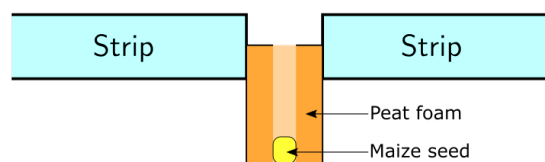
placed regularly in the bottom of each side of the tank (fig. 2.4b). The sprinklers spray nutrient solution at regular interval, set by the operator. The spraying pattern (interval and duration) can be differentiated between day and night and can be modified at any moment of the experiment. At the start of the experiment seeds (pre-germinated or not) are placed inside a foam cork and then placed inside a hole on a strip (fig 2.4c). They are 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 are available in appendix C.



(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 peatfoam cork to allow the root system to develop freely

Figure 2.4: Detailed diagrams about the phenotyping platform

After 3 days in the germination chamber, the germinated seeds were transferred onto the platform following the created design. The non-germinated seeds were discarded. Upon the start of the transfer, the sprinkler were activated and filled with Hoagland nutritive solution³ and the camera started scanning the plants. The seeds were put inside the corks with 6 ml of nutritive solution to avoid drying. During the experiment, the temperature of the greenhouse was set to 20°C at day and

³The precise concentration of the solution is presented in appendix B

18°C at night and the lights were on from 6 AM to 10 PM. The sprinklers sprayed solution 5 seconds every 295 seconds. The first tank was constantly moving but only scanning the root systems once a day. 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. All the rotation happened during 8 AM and 11 AM every morning to keep a window of 24 hours between each rotation. It frequently happened that some strips were stuck together, causing the rotation to stop and a manual intervention was needed to reset the sequence. All the interventions were done as quickly as possible to avoid any external influence on the plants.

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 XXX scales (precise to 0.01 g) for the fresh weight. They were then stored in individual envelopes and put in a stove at 70°C for 3 days to dry. After the drying process, they were weighted again, on XXX scales (precise to 0.001 g) to get the dry weight. For each plant, the remaining of the seed was consistently kept on the root system.

| retrieve scale's name

Table 2.1: 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

2.3 Data processing

Once the experiment ended, two kind of data were obtained: weight data (dry and fresh) of the fully grown plants and root scan data. Before being used as input into the two different models, the data need to be processed.

2.3.1 Weight data

For some plants, the germinated seeds placed on the platform did not continue to grow and had underdeveloped root system and sometimes no leaf system. All the plants were still weighted to avoid leaving out any data. Therefore some data points need to be removed from the dataset because they are outliers and do not represent the overall growth.

How to distinguish outliers + no too strict because the models are flexible so it's mainly to distinguish the genotypic effects correctly.

2.3.2 Root pictures

Talk about the temporal aspect of those data and how we will only use the last set of pictures in the model.

2.4 SpATS model

In the section, the SpATS model is introduced and explained. For a more thorough treatment of the model and all its components, see **rodriguez2016spatial**.

Consider a field trial of n plots arranged in a rectangular grid, where the plot positions are collected in vectors of row (r) and column (c) coordinates. If y is the vector of plot data in field order, a common model for y , to us as a starting point is

$$y = \mathbf{1}_n \beta_0 + Z_r c_r + Z_c c_c + \varepsilon \quad (2.4.01)$$

where $\mathbf{1}_n$ is a column-vector of ones of length n , c_r and c_c are, respectively, the random effect coefficients for the rows and columns and associated matrix Z_r and 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

$$y = f(u, v) + Z_r c_r + Z_c c_c + \varepsilon \quad (2.4.02)$$

where u and 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

$$f(\mathbf{u}, \mathbf{v}) = \underbrace{\mathbf{1}_n \beta_0 + \mathbf{u} \beta_1 + \mathbf{v} \beta_2 + \mathbf{u} \odot \mathbf{v} \beta_3}_{\text{Bilinear polynomial}} + \underbrace{f_u(\mathbf{u}) + f_v(\mathbf{v}) + \mathbf{u} \odot h_v(\mathbf{v}) + \mathbf{v} \odot h_u(\mathbf{u}) + f_{u,v}(\mathbf{u}, \mathbf{v})}_{\text{Smooth part}} \quad (2.4.03)$$

where \odot denotes the element-wise matrix product⁴. 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:

- $f_u(u)$ is a smooth trend along the rows, identical for all columns (i.e., a main smooth effect).
- $f_v(v)$ is a smooth trend along the columns, identical for all rows.
- $vh_u(u)$ and $uh_v(v)$ are linear-by-smooth interaction trends. For instance, $uh_v(v)$ 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, $h_v(v)$.
- $f_{u,v}(u, v)$ is a smooth-by-smooth interaction trend jointly defined over the row and column directions.

In total, six components are used to model the surface f . This may seem like a lot but this allows the translation of model 2.4.02 into a standard mixed model. a nice 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.

2.4.1 Modelling using P-splines

The functions f_u , f_v , h_u and h_v are constructed with variations 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

$$y_i = f(u_i, v_i) + \varepsilon_i, \text{ with } \varepsilon_i \sim N(0, \sigma^2). \quad (2.4.11)$$

lee_efficient_2013 show that it can be represented using B-splines. Let us form two B-splines basis:

1. one for the columns \hat{B} with $b_{il} = \hat{B}_l(u_i)$, where $\hat{B}_l(u_i)$ is the l th B-spline of the basis, evaluated at u_i
2. and one for the rows \check{B} with $b_{ip} = \check{B}_p(v_i)$, where $\check{B}_p(v_i)$ is the p th B-spline of the basis, evaluated at v_i .

⁴See Appendix A for details about the element-wise matrix product.

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

$$f(u_i, v_i) = \sum_{l=1}^L \sum_{p=1}^P \hat{B}_l(u_i) \check{B}_p(v_i) \alpha_{lp}, \quad (2.4.12)$$

where $\alpha = (\alpha_{11}, \dots, \alpha_{1P}, \dots, \alpha_{LP})^t$ is a vector of unknown regression coefficients of dimension $(LP \times 1)$. Note that \hat{B} and \check{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_curve_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,

$$B = \hat{B} \square \check{B} = (\hat{B} \otimes \mathbf{1}_L^t) \odot (\mathbf{1}_P^t \otimes \check{B}), \quad (2.4.13)$$

where the operation \square 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 (2.4.11) can be written in matrix notation

$$y = B\alpha + \epsilon. \quad (2.4.14)$$

Since the model is purely parametric, it can be estimated by minimizing the residual sum of squares (with explicit solution $\hat{\alpha} = (B^t B)^{-1} B^t y$). To prevent over-fitting, **eilers_flexible_1996** propose to incorporate a discrete penalty on the coefficient associated to adjacent B-splines. For the two-dimensional case, the vector α can be seen as an $(L \times P)$ matrix of coefficients, $A = [\alpha_{lp}]$. Now the rows and columns of A correspond to the regression coefficients in the v and u direction, respectively. In anisotropic 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 A , the other on all of its columns; and the penalized least squares objective function becomes (**eilers_multivariate_2003**)

$$S^* = \|y - B\alpha\|^2 + \hat{\lambda} \|\hat{D}A\|_F^2 + \check{\lambda} \|A\check{D}^t\|_F^2 = \|y - B\alpha\|^2 + \alpha^t P \alpha, \quad (2.4.15)$$

where $P = \hat{\lambda} (I_P \otimes \hat{D}^t \hat{D}) + \check{\lambda} (\check{D}^t \check{D} \otimes I_L)$ is the penalty matrix, $\hat{\lambda}$ and $\check{\lambda}$ are the smoothing parameters acting, respectively, on the columns and rows of A , and \hat{D} and \check{D} are the matrices that form differences of order d_u and d_v respectively⁶. The minimizer of 2.4.15 is

$$\hat{\alpha} = (B^t B + P)^{-1} B^t y. \quad (2.4.16)$$

The advantage of this decomposition is that the only tuning parameters for the smoothness of the bivariate surface are the smoothing parameters $\hat{\lambda}$ and $\check{\lambda}$.

2.4.2 Mixed model based smoothing parameters selection

As explained in **rodriguez2016spatial**, P is rank-deficient and this causes numerical instability when applying mixed model estimation techniques. To obtain a full-rank penalty matrix, the key is to write model 2.4.14 as

$$B\alpha = X_s \beta_s + Z_s c_s. \quad (2.4.21)$$

⁵See Appendix A for details about the Kronecker product

⁶See appendix A for more information about penalized splines and ordered differences

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_p-spline_2011**, based on eigenvalue decomposition which gives rise to a diagonal penalty matrix.

Let $\hat{D}^t \hat{D} = U_u E_u U_u^t$ and $\check{D}^t \check{D} = U_v E_v U_v^t$ be the eigenvalue decomposition of the marginal penalties, $\hat{D}^t \hat{D}$ and $\check{D}^t \check{D}$, respectively. Here U_j denotes the matrix of eigenvectors and E_j the diagonal matrix of eigenvalues ($j = u, v$). Let us denote \tilde{U}_j and \tilde{E}_j the sub-matrices corresponding to non-zero eigenvalues. Setting

$$\mathbf{X}_s = [\mathbf{1}_n, \mathbf{u}, \mathbf{v}, \mathbf{u} \odot \mathbf{v}] \quad \text{and} \quad \mathbf{Z}_s = [\mathbf{Z}_v, \mathbf{Z}_u, \mathbf{Z}_v \square \mathbf{u}, \mathbf{v} \square \mathbf{Z}_u, \mathbf{Z}_v \square \mathbf{Z}_u], \quad (2.4.22)$$

where $\mathbf{Z}_u = \check{B} \tilde{U}_u$ and $\mathbf{Z}_v = \check{B} \tilde{U}_v$, the penalized least problem (2.4.15) becomes

$$S^* = \|\mathbf{y} - \mathbf{X}_s \boldsymbol{\beta}_s - \mathbf{Z}_s \mathbf{c}_s\|^2 + \mathbf{c}_s^t \tilde{\mathbf{P}} \mathbf{c}_s, \quad (2.4.23)$$

where

$$\tilde{\mathbf{P}} = \text{blockdiag} \left(\check{\lambda} \tilde{\mathbf{E}}_v, \hat{\lambda} \tilde{\mathbf{E}}_u, \check{\lambda} \tilde{\mathbf{E}}_v, \hat{\lambda} \tilde{\mathbf{E}}_u, \check{\lambda} \tilde{\mathbf{E}}_v \otimes \mathbf{I}_{L-2} + \hat{\lambda} \mathbf{I}_{P-2} \otimes \tilde{\mathbf{E}}_u \right), \quad (2.4.24)$$

is the new penalty matrix. Each block of $\tilde{\mathbf{P}}$ corresponds to a block in \mathbf{Z}_s . This reformulation provides the ANOVA type decomposition discussed in the previous section (2.4.03). The block structure of \mathbf{X}_s and \mathbf{Z}_s implies

$$\begin{aligned} f(\mathbf{u}, \mathbf{v}) &= \mathbf{X}_s \boldsymbol{\beta}_s + \mathbf{Z}_s \mathbf{c}_s \\ &= \mathbf{1}_n \beta_0 + \mathbf{u} \beta_1 + \mathbf{v} \beta_2 + \mathbf{u} \odot \mathbf{v} \beta_3 \\ &\quad + \underbrace{f_v(\mathbf{v})}_{\mathbf{Z}_v \mathbf{c}_{s1}} + \underbrace{f_u(\mathbf{u})}_{\mathbf{Z}_u \mathbf{c}_{s2}} + \underbrace{\mathbf{u} \odot h_v(\mathbf{v})}_{[\mathbf{Z}_v \square \mathbf{u}] \mathbf{c}_{s3}} + \underbrace{\mathbf{v} \odot h_u(\mathbf{u})}_{[\mathbf{v} \square \mathbf{Z}_u] \mathbf{c}_{s4}} + \underbrace{f_{u,v}(\mathbf{u}, \mathbf{v})}_{[\mathbf{Z}_v \square \mathbf{Z}_u] \mathbf{c}_{s5}}, \end{aligned} \quad (2.4.25)$$

where \mathbf{c}_{sk} ($k = 1, \dots, 5$) contains the elements of \mathbf{c}_s that correspond to the k th block of \mathbf{Z}_s , i.e. $\mathbf{c}_s = (\mathbf{c}_{s1}^t, \dots, \mathbf{c}_{s5}^t)^t$.

The solution for 2.4.23, given $\check{\lambda}$ and $\hat{\lambda}$, corresponds to the best linear unbiased estimator (BLUE) for the (4×1) vector $\boldsymbol{\beta}_s$, and the BLUPs for the $((LP - 4) \times 1)$ vector \mathbf{c}_s under the linear mixed model

$$\mathbf{y} = \mathbf{X}_s \boldsymbol{\beta}_s + \mathbf{Z}_s \mathbf{c}_s + \boldsymbol{\varepsilon}, \quad \text{with } \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n) \quad \text{and} \quad \mathbf{c}_s \sim N(\mathbf{0}, \mathbf{G}_s), \quad (2.4.26)$$

where $\mathbf{G}_s = \sigma^2 \tilde{\mathbf{P}}^{-1}$. Denoting $\check{\sigma}^2 = \sigma^2 / \check{\lambda}$ and $\hat{\sigma}^2 = \sigma^2 / \hat{\lambda}$ the variance parameters involved in \mathbf{G}_s , it follows

$$\mathbf{G}_s^{-1} = \text{blockdiag} \left(\frac{1}{\check{\sigma}^2} \tilde{\mathbf{E}}_v, \frac{1}{\hat{\sigma}^2} \tilde{\mathbf{E}}_u, \frac{1}{\check{\sigma}^2} \tilde{\mathbf{E}}_v, \frac{1}{\hat{\sigma}^2} \tilde{\mathbf{E}}_u, \frac{1}{\check{\sigma}^2} \tilde{\mathbf{E}}_v \otimes \mathbf{I}_{L-2} + \frac{1}{\hat{\sigma}^2} \mathbf{I}_{P-2} \otimes \tilde{\mathbf{E}}_u \right). \quad (2.4.27)$$

Note that besides the five smooth components, \mathbf{G}_s involves only two variance parameters, $\check{\sigma}^2$ and $\hat{\sigma}^2$. In fact, 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 $\check{\sigma}^2$ and $\hat{\sigma}^2$, but in a non-linear way. Even though **rodriguez2015fast** presented a specialized algorithm to deal with this issue, here the PS-ANOVA decomposition approach

(lee_efficient_2013) is used to allow the use of standard mixed model estimation procedures. lee_efficient_2013 therefore propose to use a different variance component for each smooth component in G_s . Denoting $\Lambda_{s1}^{-1} = \Lambda_{s3}^{-1} = \tilde{E}_v$, $\Lambda_{s2}^{-1} = \Lambda_{s4}^{-1} = \tilde{E}_u$, and $\Lambda_{s5}^{-1} = \tilde{E}_v \otimes I_{L-2} + I_{P-2} \otimes \tilde{E}_u$ the precision matrix, G_s^{-1} , is then redefined as

$$G_s^{-1} = \text{blockdiag} \left(\frac{1}{\sigma_{s1}^2} \Lambda_{s1}^{-1}, \frac{1}{\sigma_{s2}^2} \Lambda_{s2}^{-1}, \frac{1}{\sigma_{s3}^2} \Lambda_{s3}^{-1}, \frac{1}{\sigma_{s4}^2} \Lambda_{s4}^{-1}, \frac{1}{\sigma_{s5}^2} \Lambda_{s5}^{-1} \right), \quad (2.4.28)$$

and thus the covariance matrix G_s is a linear function of variance parameters

$$G_s = \bigoplus_{k=1}^5 \sigma_{sk}^2 \Lambda_{sk} = \bigoplus_{k=1}^5 G_{sk} = \text{blockdiag} (G_{s1}, G_{s2}, G_{s3}, G_{s4}, G_{s5}), \quad (2.4.29)$$

where $G_{sk} = \sigma_{sk}^2 \Lambda_{sk}$ ($k = 1, \dots, 5$). 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 c_{sk} , each depending on one variance σ_{sk}^2 ($k = 1, \dots, 5$).

2.4.3 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 more complex models in order to account for the genetic variation, the presence of blocks effects, or other sources of variation (see section ??). From now on, we therefore consider the following linear mixed model

$$y = \underbrace{X_s \beta_s + Z_s c_s}_{f(u,v)} + X_d \beta_d + Z_d c_d + \varepsilon, \text{ with } c_s \sim N(0, G_s) \text{ and } c_d \sim N(0, G_d), \quad (2.4.31)$$

where X_s , Z_s and G_s are defined in the previous section and where X_d and Z_d represent column-partitioned matrices, associated respectively with extra fixed and random components, as for instance, row, column, genotypic effects. We assume that X_d has full rank, $Z_d = [Z_{d1}, \dots, Z_{db}]$ and $c_d = (c_{d1}^t, \dots, c_{db}^t)^t$. Each Z_{dk} corresponds to the design matrix of the k th random component c_{dk} , with c_{dk} being a $(m_{dk} \times 1)$ vector ($k = 1, \dots, b$). We assume further that c_s and c_d are independant, and that $G_d = \bigoplus_{k=1}^b G_{dk} = \bigoplus_{k=1}^b \sigma_{dk}^2 \Lambda_{dk}$, where Λ_{dk} are diagonal matrices. To keep the notation simple, we rewrite model (2.4.31) as

$$y = X\beta + Zc + \varepsilon, \text{ with } c \sim N(0, G) \text{ and } \varepsilon \sim N(0, \sigma^2 I_n) \quad (2.4.32)$$

where $X = [X_s, X_d]$, $Z = [Z_s, Z_d] = [Z_1, \dots, Z_q]$ ($q = 5 + b$), and

$$G = \text{blockdiag} (G_s, G_d) = \bigoplus_{k=1}^q G_k = \bigoplus_{k=1}^q \sigma_k^2 \Lambda_k \quad (2.4.33)$$

2.4.4 Model estimation

Estimates of the fixed and random effects coefficients in model **??**, for given values of the variance components, follow from standard mixed-mode theory and can be obtained by maximizing the REML log-likelihood function

$$l = -\frac{1}{2} \log |\mathbf{V}| - \frac{1}{2} \log |\mathbf{X}^t \mathbf{V}^{-1} \mathbf{X}| - \frac{1}{2} (\mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}})^t \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}}). \quad (2.4.41)$$

The implementation of this procedure in the SpATS packages is detailed in the appendix of **rodriguez2016spatial**. The REML estimates are obtained taking the derivatives of the previous equation, with respect to the variance components σ_k^2 ($k = 1, \dots, q$) and equating the obtained expression to zero (see e.g. **rodriguez2015fast**; **johnson1995restricted**), we obtain that

$$\hat{\sigma}_k^2 = \frac{\hat{\mathbf{c}}_k^t \boldsymbol{\Lambda}_k^{-1} \hat{\mathbf{c}}_k}{\text{ED}_k}, k = 1, \dots, q \quad (2.4.42)$$

with

$$\text{ED}_k = \text{trace}(\mathbf{Z}_k^t \mathbf{Q} \mathbf{Z}_k \mathbf{G}_k) \quad (2.4.43)$$

where $\mathbf{Q} = \mathbf{V}^{-1} - \mathbf{V}^{-1} \mathbf{X} (\mathbf{X}^t \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^t \mathbf{V}^{-1}$ with $\mathbf{V} = \mathbf{R} + \mathbf{Z} \mathbf{G} \mathbf{Z}^t$ and $\mathbf{R} = \sigma^2 \mathbf{I}_n$. An estimate of σ^2 can also be easily obtained following the same reasoning

$$\hat{\sigma}^2 = \frac{\hat{\boldsymbol{\varepsilon}}^t \hat{\boldsymbol{\varepsilon}}}{\text{ED}_\varepsilon} \quad (2.4.44)$$

where $\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}} + \mathbf{Z} \hat{\mathbf{c}}$ and

$$\text{ED}_\varepsilon = \text{trace}(\mathbf{R} \mathbf{Q}) = n - \text{rank}(\mathbf{X}) - \sum_{k=1}^q \text{ED}_k. \quad (2.4.45)$$

In these equations the real values of $\boldsymbol{\beta}$ and \mathbf{c} are not known, so their BLUEs ($\hat{\boldsymbol{\beta}}$) and BLUP ($\hat{\mathbf{c}}$), respectively are used instead. Their closed form expressions are detailed in the Appendix B of **rodriguez2016spatial**.

Effective dimensions

Denoting the denominator of (2.4.44) as $\text{ED}_{[\cdot]}$ (for effective dimension) was not done without purpose. In the smoothing context, the effective dimension of a "smooth" model is defined as the trace of the so-called "hat" matrix \mathbf{H} , defined as $\hat{\mathbf{y}} = \mathbf{H} \mathbf{y}$ (**hastie1990generalized**). In this setting the ED can be interpreted as a measure of the complexity of the model: the more complex the model, the larger the ED (see also **eilers2015twenty**).

In recent years, several new definitions and generalizations of the concept of effective dimension that are applicable to (generalized) linear mixed models have been proposed in the statistical literature (see, e.g., **you2016generalized**, and references therein). In almost all cases, the aim has been to provide a complexity measure that allows models' comparison and selection (via, for instance, the AIC). For our application, however, we are more interested in obtaining a separate complexity measure for each component in model **??**, that can give us insights about the contribution of that

effect when explaining the response (phenotypic) variation (**cui2010partitioning**). As shown in details in the appendix of (**rodriguez-alvarez_correcting_2018**), for model (??) we can write

$$\mathbf{H}\mathbf{y} = \hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\mathbf{c}} = \mathbf{H}_F\mathbf{y} + \mathbf{H}_R\mathbf{y} = \mathbf{H}_F\mathbf{y} + \sum_{k=1}^q \mathbf{H}_k\mathbf{y} \quad (2.4.46)$$

where \mathbf{H}_F and \mathbf{H}_R are the hat matrix for the fixed and random parts of the model, and $\mathbf{H}_k\mathbf{y}$ is the hat matrix corresponding to the random k th component \mathbf{c}_k . Since the effective dimension of \mathbf{c}_k is defined as the trace of \mathbf{H}_k then the total effective dimension of model (??) is

$$\begin{aligned} \text{ED} = \text{trace}(\mathbf{H}) &= \text{trace}(\mathbf{H}_F) + \text{trace}(\mathbf{H}_R) = \text{trace}(\mathbf{H}_F) + \sum_{k=1}^q \text{trace}(\mathbf{H}_k) \\ &= \text{rank}(\mathbf{X}) + \sum_{k=1}^q \text{ED}_k. \end{aligned} \quad (2.4.47)$$

Taking this equation and equation (2.4.45) into consideration we note that

$$n = \text{rank}(\mathbf{X}) + \sum_{k=1}^q \text{ED}_k + \text{ED}_\epsilon. \quad (2.4.48)$$

Thus, the number of observations n is partitioned into independent effective dimensions for the model's components and error.

For all components ED_k ($k = 1, \dots, q$) will vary between zero and $\text{rank}([\mathbf{X}, \mathbf{Z}_k]) - \text{rank}(\mathbf{X})$. The signal-to-noise ratio σ_k^2/σ^2 modulates the value of ED_k : when $\sigma_k^2/\sigma^2 \rightarrow 0$ then $\text{ED}_k \rightarrow 0$ and when $\sigma_k^2/\sigma^2 \rightarrow \infty$ then ED_k approaches the maximum value. ED_k can therefore be interpreted as a measure of the complexity of the corresponding component, which in turn gives a separate measure for its contribution.

For the ED_k corresponding to smoothing parameters λ_{sk} , the upper bound is determined by the number of knots used to fit the smooth surface. Therefore ED_{sk} serves as a reverse indicator of the smoothness of the corresponding components, i.e. the higher the degree of smoothness (larger value of λ_{sk}) the smaller the number of ED_{sk} , and is indicative of the relative importance of the spatial component in the construction smooth surface.

Heritability and effective dimension

The standard heritability is defined by **rodriguez-alvarez_correcting_2018** as the proportion of total (phenotypic) variation that is attributable to the genetic component. It is useful to compare the genetic effects of each genotype in a single trial but also between different models. The effective dimension associated with the genotypic effects $\text{ED}_g = \text{trace}(\mathbf{H}_g)$ is a measure of the degree of shrinkage imposed on genotypic effects, where \mathbf{H}_g is the hat matrix for genotypes. In this case, \mathbf{H}_g depends on the regularization parameter $\lambda_g = \sigma_e^2/\sigma_g^2$ and transforms the observations into predicted genotypic values, such that $\mathbf{H}_g\mathbf{y} = \mathbf{Z}_g\hat{\mathbf{g}}$. Therefore, ED_g decreases as shrinkage of

genotypic effects increases. Given the properties of the genetic effective dimension, **rodriguez2016spatial** proposed a novel expression of heritability:

$$H^2 = \frac{ED_g}{n_g - l} \quad (2.4.49)$$

where n_g is the number of genotypes and l is the number of non-zero eigenvalues of \mathbf{H}_g . Furthermore, in the specific situation when genotypic effects are assumed independent (i.e., ignoring pedigree/marker information), and by ignoring the zero eigenvalues, the following equivalence can be established:

$$H^2 = \frac{ED_g}{n_g} = 1 - \frac{\overline{PEV}}{\sigma_g^2}, \quad (2.4.410)$$

where \overline{PEV} stands for the average prediction error variance of genotype BLUPs. Note that the right-hand term corresponds to the generalized heritability developed by **welham2010compa** and is also equivalent to the one given in **cullis_design_2006**. This generalized heritability is adapted to models with additional fixed and random components in comparison to the standard heritability which only works for a model with a genotypic random component only.

Do a section explaining which model was chosen and how it was modelled in reality

2.5 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_accounting_1997**. For information about the extensions of the model, see **Piepho2010** and **williams_neighbour_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

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{c} + \boldsymbol{\xi} + \boldsymbol{\eta} \quad (2.5.01)$$

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

$$\sigma^2 \begin{bmatrix} \mathbf{G}(\boldsymbol{\gamma}) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}(\boldsymbol{\alpha}) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \psi \mathbf{I} \end{bmatrix}, \quad (2.5.02)$$

where $\psi = \sigma_{\eta}^2 / \sigma^2$, $\boldsymbol{\gamma}$ 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

$$\mathbf{y} \sim \mathbf{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 (\mathbf{Z}\mathbf{G}\mathbf{Z}^t + \mathbf{R})), \quad (2.5.03)$$

where $\mathbf{R} = \mathbf{R}(\phi) = \Sigma + \psi \mathbf{I}$, $\phi = (\alpha^t, \psi)^t$. As said in previous sections, the goal of the ARXAR model is to use a variogram to estimate the errors terms of model 2.5.01 and therefore produce better estimates of the fixed effects of the model.

2.5.1 The variogram

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

$$\omega(s, t) = \frac{1}{2} \text{var}[\mathcal{E}(s) - \mathcal{E}(t)] = \frac{1}{2}[V(s, s) + V(t, t) - 2V(s, t)], \quad (2.5.11)$$

where $s, t \in \mathbb{R}^2$ and $V(\cdot, \cdot)$ is the covariance function of $\mathcal{E}(\cdot)$. Here, $\mathcal{E}(\cdot)$ is assumed to be second-order stationary. To illustrate these concepts, we consider $e = \xi + \eta$ where e is a zero mean spatially correlated process with a directional exponential covariance (DEC) structure distributed independently of η , 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 s and t . Then

$$\omega(s, t) = \omega(\mathbf{l}) = \begin{cases} \sigma_\eta^2 + \sigma^2 [1 - \exp(-\alpha_1 l_1 - \alpha_2 l_2)] & \mathbf{l} \neq 0 \\ 0 & \mathbf{l} = 0 \end{cases}. \quad (2.5.12)$$

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 $0, d_1, 2d_1, \dots, (r-1)d_1$ and for l_2 of $0, d_2, 2d_2, \dots, (c-1)d_2$, 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 $0, 1, 2, \dots, (r-1)$ and values for l_2^* of $0, 1, 2, \dots, (c-1)$, and becomes

$$\begin{aligned} \omega(\mathbf{l}^*) &= \sigma_\eta^2 + \sigma^2 [1 - \exp(-\alpha_1 d_1 l_1^* - \alpha_2 d_2 l_2^*)] \\ &= \sigma_\eta^2 + \sigma^2 \left(1 - \rho_1^{l_1^*} \rho_2^{l_2^*}\right) & \mathbf{l}^* \neq 0 \\ &= 0 & \mathbf{l}^* = 0 \end{aligned} \quad (2.5.13)$$

where $\rho_1 = \exp(-\alpha_1 d_1)$ and $\rho_2 = \exp(-\alpha_2 d_2)$. This formulation demonstrates the equivalence between the DEC model and the AR1 x AR1 model for field experiments. Considering the model given in equation 2.5.01, the variogram ordinates for the data vector \mathbf{y} are

$$v_{ij} = \frac{1}{2} [e_i(s_i) - e_j(s_j)]^2 \quad \forall i, j = 1, \dots, n; i \neq j \quad (2.5.14)$$

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

$$\frac{v_{ij}}{\omega(s_i, s_j)} \sim \chi_1^2 \quad (2.5.15)$$

so that v_{ij} is unbiased for $\omega(s_i, s_j)$. 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 $(l_{ij1}, l_{ij2}, \bar{v}_{ij})$, where $l_{ij1} = |s_{i1} - s_{j1}|$ and $l_{ij2} = |s_{i2} - s_{j2}|$ are the absolute displacements and \bar{v}_{ij} is the sample mean of the v_{ij} with the same absolute displacements.

2.5.2 Model estimation

The result that v_{ij} is unbiased for $\omega(s_i, s_j)$ is based on the assumptions that β and c are known. In practice, we replace β and c by their GLS estimates ($\hat{\beta}$) and BLUP (\hat{c}) respectively, so that the BLUP of the residual vector is given by

$$\hat{e} = y - X\hat{\beta} - Z\hat{c} = RP_y \quad (2.5.21)$$

where $P = R^{-1} - R^{-1}WC^{-1}W^tR^{-1}$ with $W = [XZ]$ and $C = W^tR^{-1}W + G^*$ is the coefficient matrix from the mixed model equation and it is partitioned in the same way as W . G^* is a square matrix of order $t + b$, partitioned similarly to $W^tR^{-1}W$ and is 0 except in the lower diagonal block corresponding to $Z^tR^{-1}Z$, where it equals G^{-1} .

Under the assumption of a Gaussian distribution for y , $\hat{e} \sim N(0, \sigma^2(R - WC^{-1}W^t))$ assuming (γ, ϕ) is known. Following the decomposition of (2.5.14), variogram ordinates v_{ij} can be expressed as a quadratic form in y , that is

$$v_{ij} = (a_{ij}^t e) (a_{ij}^t e) = e^t a_{ij} a_{ij}^t e = e^t A_{ij} e \quad (2.5.22)$$

and similarly

$$\hat{v}_{ij} = (a_{ij}^t \hat{e}) (a_{ij}^t \hat{e}) = \hat{e}^t a_{ij} a_{ij}^t \hat{e} = \hat{e}^t A_{ij} \hat{e} \quad (2.5.23)$$

where $A_{ij}^{(n \times n)}$ 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

$$\begin{aligned} E(\hat{v}_{ij}) &= \sigma^2 \text{trace} [A_{ij} (R - WC^{-1}W^t)] \\ &= \sigma^2 \text{trace} [A_{ij} R] - \sigma^2 \text{trace} [A_{ij} WC^{-1}W^t] \\ &= \sigma^2 a_{ij}^t R a_{ij} - \sigma^2 a_{ij}^t W C^{-1} W^t a_{ij} \\ &= E(v_{ij}) - \sigma^2 a_{ij}^t W C^{-1} W^t a_{ij} \end{aligned} \quad (2.5.24)$$

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

$$WC^{-1}W^t = \sum_{k=1}^{t+b} \lambda_k w_k w_k', \quad (2.5.25)$$

then

$$\begin{aligned} E(\hat{v}_{ij}) &= E(v_{ij}) - \sigma^2 a_{ij}^t \left(\sum_k \lambda_k w_k w_k' \right) a_{ij} \\ &= E(v_{ij}) - \sigma^2 \sum_h \lambda_h (a_{ij}^t w_h)^2 \\ &= E(v_{ij}) - \sigma^2 \sum_k \lambda_k w_k' A_{ij} w_k \end{aligned} \quad (2.5.26)$$

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 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 2.5.26 is approximate. The effect of the estimation of (γ, ϕ) on the distribution of \hat{e} (and functions of \hat{e}) is an important problem. **kenward1997precision** have examined this issue for the testing of fixed effects in REML.

2.5.3 Base model selection

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

2.5.4 Extension to the linear variance (LV) model

| Explain how the model were fitted

2.6 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_modelling_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_design_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 2.4.410 to perform a straightforward comparison between the heritability estimated by SpATS and that obtained from the BSS model (**velazco_modelling_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 **rodriguez-alvarez_correcting_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 3

Results and discussion

3.1 Pre-processing and first analysis

3.2 SpATS analysis

3.3 ARxAR model analysis

3.4 Model comparison

3.4.1 Performances

3.4.2 Parametrization

3.4.3 Modelling strategy

Chapter 4

Conclusion

Appendices

Appendix A

Additional informations on computation

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

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:

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

Polynomials splines

fahrmeir_regression: 2013 state that a function $f : [a, b] \rightarrow \mathbb{R}$ is called a polynomial spline of degree $l \geq 0$ with knots $a = \kappa_1 < \dots < \kappa_m = b$, 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 $\kappa_1 < \dots < \kappa_m$ can be uniquely determined as a linear combination of the $d = l + m - 1$ functions B_1, \dots, B_d , called the *basis functions*, since we can uniquely represent all polynomials splines by using these functions.

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

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

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

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

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.

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

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

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 r th-order derivatives. This leads to the penalized residual sum of squares

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

where Δ^r denotes the r th-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

$$\text{PLS}(\lambda) = (\mathbf{y} - \mathbf{Z}\boldsymbol{\gamma})'(\mathbf{y} - \mathbf{Z}\boldsymbol{\gamma}) + \lambda \boldsymbol{\gamma}' \mathbf{K}_r \boldsymbol{\gamma}$$

where \mathbf{K}_r is the r th-order difference penalty matrix, and can be decomposed as $\mathbf{D}_r' \mathbf{D}_r$ with \mathbf{D}_r the r th-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

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

For more detailed information about polynomials splines, please refer to **fahrmeir_regression:_2013** and **eilers_flexible_1996**

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 ₂ O	472	62.5
2M MgSO ₄ x 7 H ₂ O	493	25
1M NH ₄ NO ₃	80	25
Minors:		
H ₃ BO ₃	2.86	
MnCl ₂ x 4 H ₂ O	1.81	
ZnSO ₄ x 7 H ₂ O	0.22	25 ²
CuSO ₄	0.051	
H ₃ MoO ₄ x H ₂ O or	0.09	
Na ₂ MoO ₄ x 2 H ₂ O	0.12	
1M KH ₂ PO 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

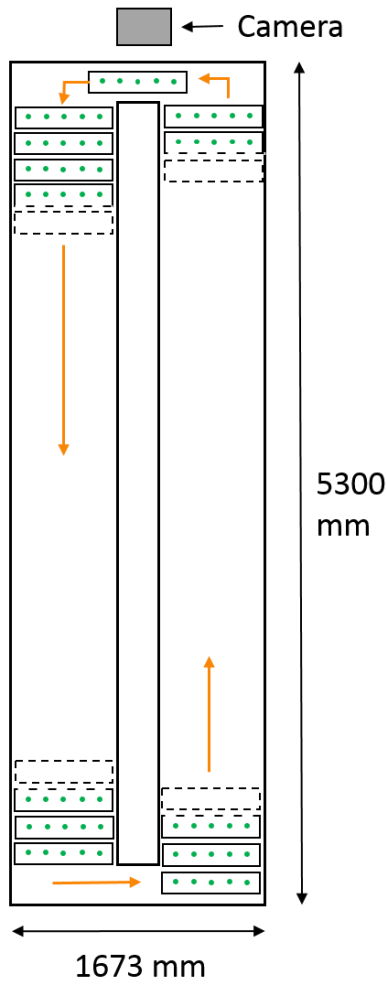
JRA2 - Jan. 2018



Platform name

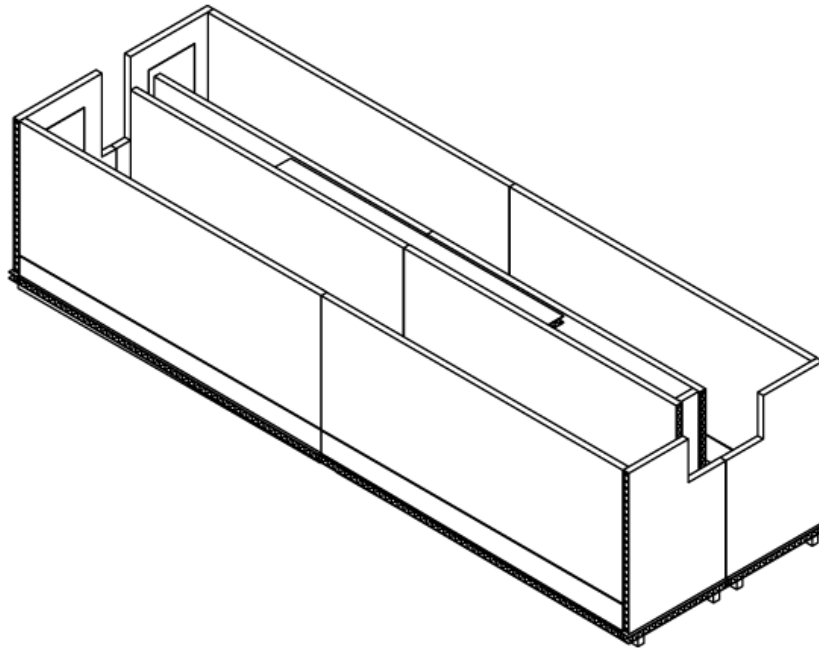
Partner site	UCL
Site and installation	Site: Louvain-la-Neuve, Installation: Aeroponics
Contact person(s)	Xavier Draye xavier.draye@uclouvain.be

Description of the platform structure

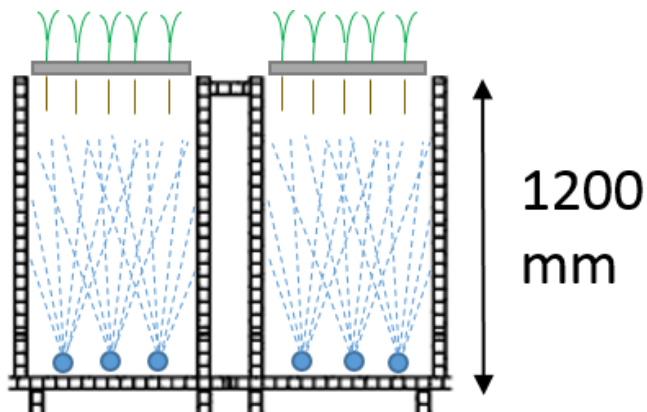


JRA2 - Jan. 2018

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.



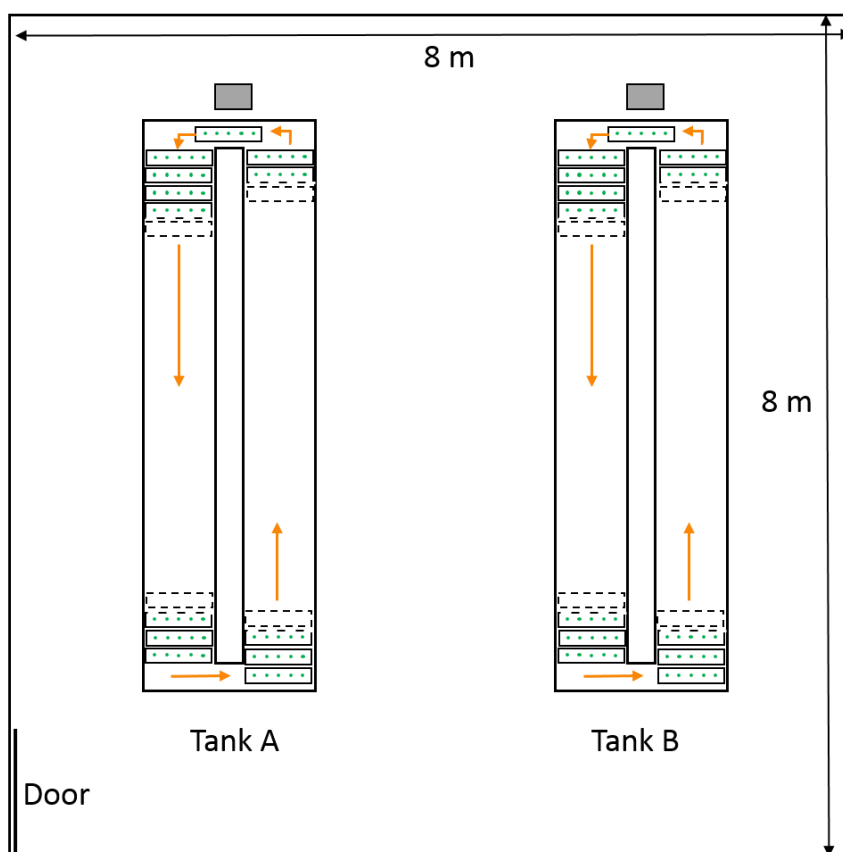
Transversal view of the aeroponic tank: 3 sprinklers are placed regularly in the bottom of each side of the tank. The sprinklers spray nutrient solution at regular interval, set by the operator. The spraying



JRA2 - Jan. 2018

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.

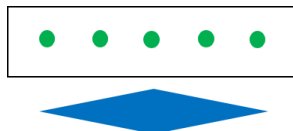


Sources and directions (if known) of environmental variations in the installation

- 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.

JRA2 - Jan. 2018

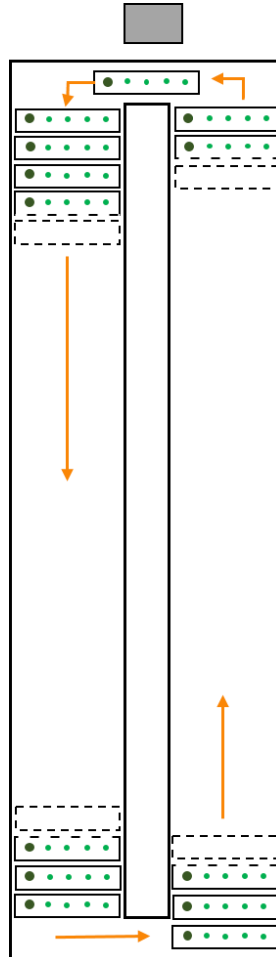
- 3) Inside each tank, between plants that grow in the middle of the strip as compared to plants growing at the border of the strip. We suppose that the plants at the extremity of the strip may receive a bit less water than the others.



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.

JRA2 - Jan. 2018



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

JRA2 - Jan. 2018



- 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:

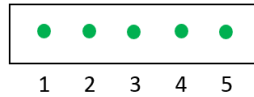
Ex: B_76_5

B: tank id (A or B)

76: strip id (from 1 to 99)

5: position in the strip (from 1 to 5)

JRA2 - Jan. 2018



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>



