eman ta zabal zazu

Universidad del País Vasco

Euskal Herriko Unibertsitatea

Spatio-temporal modelling of high-throughput phenotyping data

Diana Marcela Pérez Valencia

Supervised by María Xosé Rodríguez Álvarez and Fred A. van Eeuwijk

October 2023

Spatio-temporal modelling of high-throughput phenotyping data

Diana Marcela Pérez Valencia

Supervised by María Xosé Rodríguez Álvarez and Fred A. van Eeuwijk

October 2023

This research was supported by the project MTM2017-82379-R (AEI/FEDER, UE), the Basque Government through the BERC 2018-2021 and BERC 2022-2025 programs, and by the Spanish Ministry of Science and Innovation: BCAM Severo Ochoa accreditation SEV-2017-0718 and BCAM Severo Ochoa accreditation CEX2021-001142-S/MICIN/AEI/10.13039/501100011033. We specially thank Llorenç Cabrera-Bosquet and François Tardieu (LEPSE, INRAE, Montpellier, France) for sharing with us the PhenoArch data, Lukas Kronenberg and Andreas Hund (ETH Zürich, Switzerlad) for sharing with us the FIP data, and Martin P. Boer, Emilie J. Millet, Bart-Jan van Rossum and Daniela Bustos-Korts for the discussions and support on my research stays at Biometris, Wageningen University & Research (The Netherlands).

Acknowledgements

A mis asesores, por apoyarme en cada detalle. A Coté, por enseñarme tanto y exigirme cada vez más, cada vez mejor. A Fred, por ver más allá, por su amplia perspectiva de las cosas.

A todos los investigadores que de alguna manera han revisado el material de esta tesis por sus valiosos aportes que han contribuodo a mejorar este trabajo.

Al equipo de "P-splines fiesta" y a las personas involucradas en el curso de "Plant Genetics, Genomics and Breeding" en Zaragoza por motivarme con tantos conocimientos.

A mi tutora Inmaculada Arostegui por siempre estar al tanto de mi proceso y a mi mentora, Anabel Forte, por todos sus generosos consejos.

Al personal del BCAM y de la UPV/EHU por el apoyo en cada etapa.

Al grupo de investigación de Estadística Aplicada del BCAM y al grupo de investigación Biometris de la Universidad de Wageningen por acogerme tan calurosamente.

A tantos amigos que he tenido la fortuna de conocer en estos cinco años por ser mi familia extranjera, por inspirarme con su energía y juventud y por compartirme la belleza de sus culturas.

A mis "viejos" amigos por ser siempre una parte importante de mi vida. Especialmente a aquel que siempre me anima a superar mis límites.

A mi familia, y en especial a mi mamá, mi papá y mi hermanita, por su amor incondicional y por disfrutar conmigo de este viaje.

A Bilbao por ser mi segundo hogar después de mi muy amado Medellín-Medellín.

Abstract

High throughput phenotyping (HTP) platforms and devices are increasingly used to characterise growth and developmental processes for large sets of plant genotypes. This dissertation is motivated by the need to accurately estimate genetic effects over time when analysing data from such HTP experiments. The HTP data we deal with here are characterised by phenotypic traits measured multiple times in the presence of spatial and temporal noise and a hierarchical organisation at three levels (populations, genotypes within populations, and plants within genotypes). The challenge is to balance efficient statistical models and computational solutions to deal with the complexity and dimensionality of the experimental data. To that aim, we propose two strategies. The first proposal divides the problem into two stages. The first stage (spatial model) focuses on correcting the phenotypic data for experimental design factors and spatial variation, while the second stage (hierarchical longitudinal model) aims to estimate the evolution over time of the genetic signal. The second proposal is to face the problem simultaneously (one-stage approach). That is, modelling the longitudinal evolution of the genetic effect on a given phenotypic trait while accounting for the temporal and spatial effects of environmental and design factors (spatio-temporal hierarchical model). We follow the same modelling philosophy throughout our work and propose multidimensional P-spline-based hierarchical approaches. We provide the user with appealing tools that take advantage of the sparse model matrices structure to reduce computational complexity. All our codes are publicly available on the R-package `statgenHTP` and https://gitlab.bcamath.org/dperez/htp_one_stage_approach. We illustrate the performance of our methods using spatio-temporal simulated data and data from the PhenoArch greenhouse platform at INRAE Montpellier and the outdoor Field Phenotyping platform at ETH Zürich. In the plant breeding context, we show how to extract new time-independent phenotypes for genomic selection purposes.

Resumen

El uso de técnicas y plataformas de fenotipado de alto rendimiento (HTP por sus siglas en inglés, "high-throughput phenotyping") se ha incrementado significativamente en los últimos años en aplicaciones de genética y fisiología de plantas. Los últimos avances en plataformas HTP han sido revisados en estudios recientes (Jin et al., 2020; D. Li et al., 2021; Song et al., 2021; Q. Xiao et al., 2022; W. Yang et al., 2020). Por ejemplo, D. Li et al. (2021) presentan una revisión de estas plataformas a nivel mundial: 12 en invernaderos y 34 en campo abierto (18 de ellas se midieron con sensores terrestres y 16 con sensores aéreos de gran escala). Los datos de HTP proporcionan información rápida, precisa, no-destructiva y costo-eficiente sobre rasgos fenotípicos con una alta resolución temporal y espacial (Tardieu et al., 2017). El diseño de experimentos en HTP, tanto en invernadero como en campo, generalmente consiste en unidades experimentales (p.e., plantas individuales en macetas o parcelas) que se combinan con una amplia gama de sensores para hacer un seguimiento (casi) continuo de los rasgos fenotípicos para grandes conjuntos de plantas y genotipos. Los datos de HTP, se obtienen de múltiples sensores (p.e., imágenes, nubes de puntos, datos hiperespectrales) y son normalmente filtrados, condensados, integrados y resumidos en características. Combinaciones de una o varias características se usan para aproximar rasgos biológicos que, por un lado, siguen estando próximos a los datos y, por otro, están relativamente alejados de los rasgos objetivo de interés comercial (la mayoría de las veces, parámetros de rendimiento y calidad). Ejemplos de estos rasgos son la altura de la planta, la cobertura del dosel, el índice de área foliar, el recuento de espigas y tallos, la temperatura de la copa, o los índices relacionados con el contenido de agua o clorofila. Por ejemplo, uno de los experimentos más grandes de HTP en invernadero (en términos de capacidad del invernadero) es el desarrollado por W. Yang et al. (2014) para 533 genotipos de *O. sativa landrace* y *elite* en 5472 plantas de arroz. Para este experimento, los autores extrajeron 15 rasgos diferentes. Aunque no se dispone de información sobre el tiempo transcurrido entre las mediciones, este experimento es evidencia de la enorme cantidad de información (datos) que producen estas plataformas.

Para facilitar el progreso en el fenotipado de plantas, la inversión y la colaboración siguen siendo esenciales. Diversas organizaciones, universidades e iniciativas de todo el mundo contribuyen en este ámbito. En particular, la Iniciativa de la Asociación Mundial para el Fomento de la Capacidad de Mejora Vegetal (GIPB), convocada por la Organización de las Naciones Unidas para la Agricultura y la Alimentación (FAO),

ejemplifica el compromiso con la colaboración mundial para la mejora y la difusión eficaz de las variedades de cultivos. Adicionalmente, se han conformado diferentes redes que proporcionan plataformas HTP para la colaboración, el intercambio de conocimientos y la creación de capacidad en materia de fenotipado de plantas. Aunque la principal desventaja de las plataformas HTP es su alto costo, su ventaja radica en que permiten controlar, cuantificar y evaluar continuamente fenotipos específicos para experimentos agrícolas de gran escala con alta resolución y precisión. Por consiguiente, con estas plataformas HTP, investigadores y fitomejoradores tienen acceso a grandes (y detallados) conjuntos de datos, en forma de (largas) series temporales, que permiten seguir múltiples rasgos biológicos desde, por ejemplo, la germinación de la semilla hasta su madurez fisiológica. Independientemente de si se miden rasgos complejos (como el rendimiento) una sola vez (como tradicionalmente) o si se miden rasgos fenotípicos varias veces (como en HTP), se sabe que su expresión se ve afectada espacialmente por factores ambientales como la heterogeneidad del suelo en los experimentos de campo y por gradientes de temperatura y luz en experimentos de invernadero. Esto hace necesario corregir estos factores de ruido cuando se analizan experimentos agrícolas (Araus & Cairns, 2014; van Eeuwijk et al., 2019). Sin embargo, los datos de los experimentos HTP presentan, además de la variación espacial, una dimensión temporal que debe incorporarse y modelarse adecuadamente en los análisis.

Consideramos que, en los experimentos de HTP, los genotipos se han asignado a las unidades experimentales siguiendo un diseño experimental. En particular, los datos HTP que tratamos en esta tesis se caracterizan por rasgos fenotípicos medidos múltiples veces en presencia de ruido espacial y temporal. Para simplificar, en lo que sigue suponemos que sólo se han analizado los genotipos (es decir, tenemos una estructura de tratamiento de un solo factor). Sin embargo, permitimos una estructura de población, modelada como diferentes familias, paneles o poblaciones de genotipos. Así, los datos presentan una estructura jerárquica anidada de tres niveles, con plantas/parcelas anidadas en genotipos, y genotipos anidados en poblaciones. Para mayor claridad y simplicidad, en lo sucesivo nos referiremos a las unidades experimentales del diseño experimental como plantas. La unidad experimental suele ser una planta en un experimento de invernaderos y una parcela con varias plantas en un experimento de campo. **El principal objetivo de esta tesis doctoral es centrarse en la estimación (y posterior procesamiento y análisis) de la evolución temporal del efecto genético sobre un fenotipo específico, al tiempo que se corrigen efectos de ruido ambientales tanto espacial como temporalmente. Al abordar este objetivo, esta investigación se enfrenta al gran desafío de combinar métodos estadísticos y computacionales que exploten eficaz y adecuadamente la diversidad y complejidad de los datos de HTP** para: (a) extraer información relevante relacionada con el crecimiento y desarrollo de las plantas, (b) incrementar la comprensión biológica de los sistemas vegetales, y (c) apoyar el proceso de toma de decisiones en los programas de fitomejoramiento. Las técnicas estadísticas desarrolladas para hacer frente a los retos que plantean los datos de HTP serán de gran interés no sólo para los fitomejoradores y estadísticos de ese campo, sino también para los profesionales que trabajan en medicina, genética humana y animal, biología evolutiva y otros ámbitos. Esta tesis requirió una sólida base

en estadística, métodos computacionales y conocimientos en sistemas vegetales. Implicó el desarrollo e implementación de modelos espacio-temporales, el diseño y desarrollo de software, y la evaluación y apoyo con datos reales y simulados. El objetivo último de esta investigación es contribuir en el mejoramiento de las prácticas de fitomejoramiento y, en consecuencia, de la seguridad alimentaria mundial.

La revisión de la literatura en modelos espacio-temporales para datos HTP (ver Capítulo 1) ha revelado que ésta es un área de investigación en auge. Muestra de esto es que agricultores y agrónomos buscan optimizar el rendimiento de los cultivos y mejorar las prácticas agrícolas mediante la toma de decisiones basada en datos. A continuación exponemos algunos de los puntos metodológicos más relevantes que abordamos en esta disertación

(a) Decidir el número de etapas (es decir, si utilizar un enfoque por etapas o un enfoque de una sola etapa) y el orden de las etapas (qué componente modelar primero, el espacial o el temporal) para los enfoques propuestos.

(b) Seleccionar y combinar métodos estadísticos apropiados para el análisis de cada componente de los datos (espacial y temporal).

(c) Aprovechar la estructura de correlación espacio-temporal y jerárquica de los datos para evitar la pérdida de información (cuando sea posible/necesario) entre y dentro de las etapas, mejorar la precisión y robustez de los modelos estadísticos utilizados, y predecir mejor los datos.

(d) Combinar métodos estadísticos y computacionales para hacer frente al tamaño y la complejidad/dimensionalidad de los datos. Esto requiere algoritmos eficientes y el desarrollo de software que pueda manejar grandes conjuntos de datos

(e) Analizar datos HTP reales. La propia recopilación y análisis de datos HTP plantea varios desafíos: el diseño del experimento, la recopilación de los datos, el procesamiento de imágenes y la limpieza de los datos. Estos son procesos que pueden requerir mucho tiempo y recursos, además de ser costosos. La inversión en la adquisición de datos HTP es una prueba de la relevancia de los proyectos en este campo. Sin embargo, el alcance de este proyecto asume que los datos ya han sido recopilados y están listos para ser analizados.

(f) Reproducir lo más fielmente posible la dinámica del sistema real utilizando datos simulados como alternativa/oportunidad para evaluar los métodos propuestos en un conjunto de escenarios y configuraciones diferentes. (Bustos-Korts et al., 2019).

(g) Desarrollar software escalable y fácil de usar para los datos HTP y hacerlo accesible a la comunidad científica.

Para dar cumplimiento al objetivo de esta investigación, hemos organizado la tesis como sigue. En el Capítulo 2 motivamos esta propuesta presentando dos conjuntos de datos de HTP. Se trata de la plataforma PhenoArch de un invernadero en INRAE Montpelier y la plataforma FIeld Phenotyping (FIP) en ETH Zürich. En este capítulo, hacemos un análisis estadístico descriptivo para comentar sobre diferentes características de los datos como: patrón de datos faltantes, forma de las curvas, comportamiento de la variabilidad entre e intra curvas, posibles interacciones entre factores, y la variación espacial (y cómo cambia a lo largo del tiempo). En el caso de la plataforma FIP, se dispone de tres ensayos independientes (tres años: 2015, 2016 y 2017). Por lo tanto, para estos datos en particular, comentamos además sobre la consistencia del desempeño genotípico a lo largo de los ensayos.

El Capítulo 3 contiene el soporte teórico de esta tesis con conceptos sobre P-splines y su extensión al caso multidimensional mediante el uso de productos tensoriales. Se hace uso de dos dimensiones para los modelos espaciales (en las direcciones de fila y columna) y de tres dimensiones para los modelos espacio-temporales (en las direcciones de fila, columna y tiempo). Explicamos con más detalle la conexión entre P-splines y modelos mixtos como marco de estimación. En este capítulo se esbozan algunos detalles técnicos esenciales utilizados principalmente en los Capítulos 4 y 5.

Las principales contribuciones metodológicas de esta tesis se presentan en los Capítulos 4 y 5. Para cumplir el objetivo principal de esta tesis, hemos propuesto dos aproximaciones al mismo problema. La primera propuesta divide el problema en dos etapas (Capítulo 4). La primera etapa (modelo espacial) se centra en corregir los datos fenotípicos para los factores de diseño experimental y la variación espacial, mientras que la segunda etapa (modelo longitudinal jerárquico) tiene como objetivo estimar la evolución en el tiempo de la señal genética. La segunda propuesta consiste en afrontar el problema simultáneamente (enfoque de una etapa, Capítulo 5). Es decir, modelizar la evolución longitudinal del efecto genético sobre un determinado rasgo fenotípico al tiempo que se tienen en cuenta los efectos temporales y espaciales de los factores ambientales y de diseño (modelo espacio-temporal jerárquico). Seguimos la misma filosofía de modelización en todo nuestro trabajo y proponemos modelos basados en P-splines.

Para ambos enfoques, explotamos la conexión entre P-splines y modelos lineales mixtos y proponemos utilizar herramientas computacionalmente atractivas que aprovechan la estructura dispersa de las matrices implicadas en los modelos para reducir la complejidad computacional. Como resultado, obtenemos curvas estimadas y sus derivadas en los tres niveles de la jerarquía (poblaciones, genotipos y plantas). Utilizamos estas curvas para extraer diferentes características independientes del tiempo. Estas características pueden servir como entradas para análisis estadísticos posteriores que tengan como objetivo modelar las interacciones genotipo-ambiente en rasgos biológicamente complejos, como el rendimiento, en relación con componentes de rasgos subyacentes (véase, por ejemplo, Moreira et al., 2020; van Eeuwijk et al., 2019).

Los Capítulos 6 y 7 están dedicados a comparar y evaluar el rendimiento de los enfoques propuestos en una y dos etapas. En el Capítulo 6, se usan datos espacio-temporales simulados, mientras que en el

Capítulo 7, analizamos los dos conjuntos de datos de HTP descritos en el Capítulo 2. Para garantizar la consistencia entre los dos enfoques, establecemos un entorno común de comparación tanto para la simulación como para el análisis de datos reales. El estudio de simulación pretende respaldar los resultados de nuestra investigación y validar nuestros enfoques explorando el impacto de diferentes variables y escenarios en un entorno controlado. Los estudios de simulación en este entorno son poco frecuentes (véase, por ejemplo, Bustos-Korts et al., 2019; Roth et al., 2021), en parte debido a la complejidad (estadística y biológica) del sistema, que hace que el mecanismo de generación de datos sea una tarea difícil. Basamos el escenario de simulación y el modelo de generación de datos en nuestros aprendizajes con análisis de datos de HTP reales. Para minimizar el sesgo, proponemos un modelo de generación de datos que es independiente de los modelos estadísticos utilizados para su análisis posterior. Somos conscientes de que nuestras simulaciones implican algunas simplificaciones y suposiciones que pueden no reflejar con exactitud las complejidades y características específicas de los datos de HTP reales. Por ello, para contrastar los resultados de la simulación, analizamos los datos de dos plataformas HTP diferentes. Los resultados obtenidos con estos dos ejemplos proporcionan información sobre factores no considerados en la simulación, que constituirán importantes líneas de investigación futura.

En un esfuerzo por compartir nuestro trabajo y contribuir al progreso del conocimiento científico, hemos implementado los enfoques propuestos en el lenguaje `R` (R Core Team, 2023). Para los enfoques de una y dos etapas, proponemos funciones para ajustar, predecir, graficar y extraer características independientes del tiempo. Mostramos las funcionalidades del código disponible en el Capítulo 8. Al compartir nuestro código, esperamos garantizar la reproducibilidad de nuestros resultados y promover la colaboración en la comunidad científica. Proporcionamos al usuario un ejemplo para reproducir los resultados de la plataforma PhenoArch. En consecuencia, es nuestro deseo que otros investigadores puedan verificar nuestros hallazgos, reproducir el análisis y tener como referencia nuestro trabajo. Además, pretendemos acercar a los científicos de las plantas a herramientas estadísticas fáciles de usar para apoyar su proceso de toma de decisiones.

La mayor parte del contenido de esta tesis ha sido discutido en diferentes espacios académicos. El Capítulo 4 y parte del Capítulo 7 fueron publicados en Scientific Reports (Perez-Valencia et al., 2022). Los Capítulos 5, 6 y 7 fueron sometidos al Journal of Agricultural, Biological and Environmental Statistics (JABES), pero está disponible en su versión BioXiv (Perez-Valencia et al., 2023). Todas las funciones presentadas en el Capítulo 8 están disponibles públicamente a través del paquete de `R statgenHTP` (enfoque en dos etapas) y en https://gitlab.bcamath.org/dperez/htp_one_stage_approach (enfoque en una etapa). Finalmente, esta tesis termina con unas conclusiones en el Capítulo 9, en donde se resumen las contribuciones de esta tesis y se discuten algunas líneas de investigación futuras. En resumen, creemos que esta tesis representa un punto de partida prometedor para el análisis espacio-temporal de datos de HTP jerárquicos. Los dos enfoques propuestos representan un buen compromiso entre flexibilidad, precisión, adecuación, eficiencia computacional e interpretabilidad. Nuestros resultados demuestran la viabilidad de nuestras propuestas en ordenadores estándar, proporcionando valiosas descripciones de la variación genética (y no genética) en la

dimensión temporal y estadísticas de resumen útiles con fines de selección genotípica. Creemos que nuestra propuesta representa una herramienta poderosa para su aplicación rutinaria en experimentos de fenotipado con series temporales densas.

Contents

Chapter 1

Introduction

1.1 Context: High-throughput phenotyping (HTP) platforms

The use of high-throughput phenotyping (HTP) techniques and platforms has significantly increased in recent years in plant genetics and physiology. The latest advancements in HTP platforms have been reviewed in recent studies (Jin et al., 2020; D. Li et al., 2021; Song et al., 2021; Q. Xiao et al., 2022; W. Yang et al., 2020). For instance, D. Li et al. (2021) provide an overview on HTP platforms worldwide, both indoors (a total of 12) and in a field (a total of 18 for ground-based proximal phenotyping, and 16 for aerial large-scale remote sensing). HTP data provide quick, precise, non-destructive and cost-effective information on phenotypic traits with high spatial and temporal resolution (Tardieu et al., 2017). Designed HTP experiments, either indoors or in a field, usually consist of experimental units (e.g., single plants in pots or plots) that are combined with a wide range of sensing equipment for the (almost) continuous monitoring of plant/plot phenotypic traits for large sets of genotypes. High dimensional HTP data as derived from multiple sensors (e.g., images, point clouds, hyperspectral data) are typically filtered, condensed, integrated and summarised into features. Combinations of one or more features are used to approximate biological traits that are on the one hand still close to the data, and on the other hand are relatively far from the target traits of commercial interest (most often yield and quality parameters). Examples of such traits are plant height, canopy cover, leaf area index, ear and tiller counts, canopy temperature or indices related to water or chlorophyll content. As an illustration, one of the largest (in terms of greenhouse capacity) indoor HTP experiment is the one developed by W. Yang et al. (2014) for 533 *O. sativa landrace* and *elite* genotypes on 5472 rice plants. For that experiment, the authors extracted 15 different traits. Although information on the time between measurements is unavailable, this experiment highlights the enormous amount of information (data) produced by these platforms.

To facilitate progress in plant phenotyping, investment and collaboration remain essential. Various or-

ganisations, universities, and initiatives worldwide contribute to this domain (D. Li et al., 2021). Notably, the Global Partnership Initiative for Plant Breeding Capacity Building (GIPB), convened by the Food and Agriculture Organisation (FAO), exemplifies the commitment to global collaboration for the effective improvement and dissemination of crop varieties. However, networking is required to provide platforms for collaboration, knowledge sharing, and capacity building in plant phenotyping. With these aims in mind, several networks have emerged, including the International Plant Phenotyping Network (IPPN), Global Plant Phenotyping Network (GPPN), International Crop Phenotyping Initiative (ICPI), European Plant Phenotyping Network (EPPN), African Plant Phenotyping Network (APPN), Phenotyping Network of the Americas (PanPheno), Brazilian Phenotyping Network (BraPhen), Argentine Plant Phenotyping Network (RFAFaneg), Australian Plant Phenomics Facility (APPF), and the Plant Accelerator (Australia).

Even though the main disadvantage of HTP platforms is that they are expensive, as noted before the advantage relies on continuously monitoring, quantifying, and evaluating specific phenotypes for large-scale agricultural experiments with high resolution and precision. Consequently, with these HTP platforms, researchers and plant breeders have now access to large and detailed datasets, in the form of (long) time-series, enabling to track multiple biological traits from, e.g., seed emergence to physiological maturity. Regardless of the fact that endpoint traits (like yield) are measured only once (as traditionally) or phenotypic traits are measured several times (as in HTP), it is known that their expression is spatially affected by environmental factors such as soil heterogeneity in field experiments and temperature and light gradients in the greenhouse. This makes it necessary to correct for these nuisance factors when analysing agricultural experiments (Araus & Cairns, 2014; van Eeuwijk et al., 2019). Yet, data from HTP experiments present, on top of spatial variation, a time dimension which needs to be incorporated and adequately modelled in the analyses.

Therefore, the main objective of this PhD thesis is to focus on estimating (and further processing and analysing) the temporal evolution of the genetic effect on a specific phenotype, while correcting for environmental nuisance effects both spatially and temporally. By addressing this objective, this research contributes to the advancement of spatio-temporal modelling methods in plant breeding and software development from a statistical perspective. The statistical techniques developed to address the challenges of HTP data will be of significant interest not only to plant breeders and statisticians in that field but also to professionals working in medicine, human and animal genetics, evolutionary biology, and beyond.

1.1.1 Spatio-temporal and hierarchical HTP data structure

We consider that, in the HTP experiment, genotypes have been allocated to the experimental units following an experimental design. Particularly, the HTP data we deal with in this thesis (see Figure 1.1) are characterised by phenotypic traits measured multiple times in the presence of spatial and temporal noise. For simplicity, in what follows we assume that only genotypes have been tested (i.e., we have a single factor

treatment structure). However, we allow for population structure, modelled as different families, panels or populations of genotypes. Thus, the data present a three-level nested hierarchical structure, with plants/plots nested in genotypes, and genotypes nested in populations. For clarity and simplicity, hereafter we refer to the experimental units of the experimental design as plants. The experimental unit is most commonly a plant for an indoor experiment and a plot containing several plants for a field experiment.

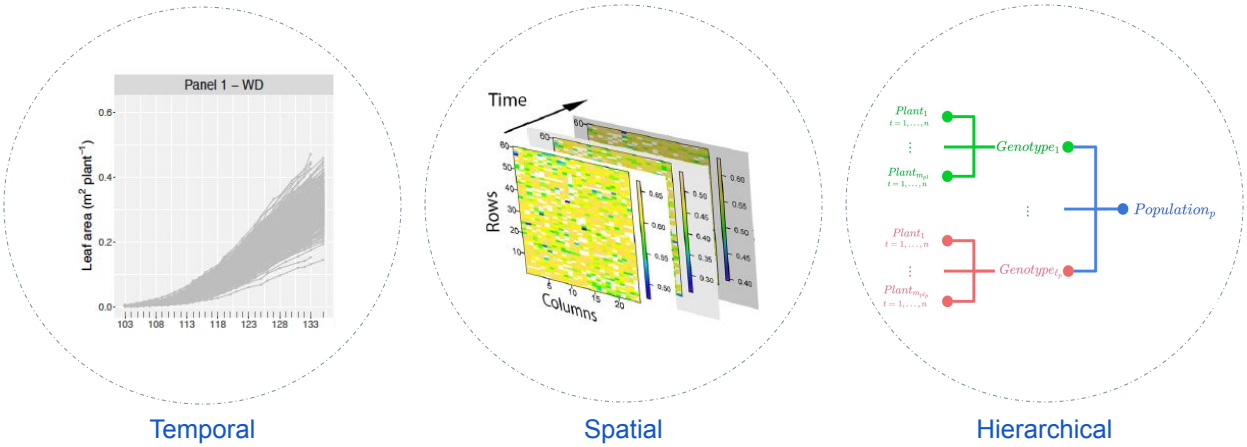


Figure 1.1: Spatio-temporal and three-level hierarchical HTP data structure.

In mathematical notation, let $y_i(t)$ denote the observed phenotypic trait of interest for the ith plant ($i = 1, \ldots, M$) at time, $t \in \{t_1, \ldots, t_n\}$. We use p and g for indices of population and genotype, respectively ($p = 1, \ldots, K$; $g = 1, \ldots, L$). With a slight abuse of notation we use $p(i)$ and $g(i)$ to denote the population and genotype of the ith plant, respectively. Let $\ell_p = \#\{g \mid p(g) = p\}$ be the number of genotypes with population p, and $m_g = \#\{i \mid g(i) = g\}$ be the number of plants with genotype g. Consequently, $L = \sum_{p=1}^{K} \ell_p$ denotes the total number of genotypes and $M = \sum_{g=1}^{L} m_g$ the total number of plants. We consider that plants can be mapped to a coordinate system defined in terms of R rows and C columns, and denote as $r(i)$ and $c(i)$ the row and column position respectively of the ith plant ($r = 1, \ldots, R$; $c = 1, \ldots, C$). We assume that all plants in the experiment are measured at the same times ($\{t_1, \ldots, t_n\}$). That is a simplification, as platform data is typically acquired within the order of minutes to hours. However, we presume that the factors that may affect the platform measurements within that period can be accounted for (and captured) by the experimental design (e.g., blocking structure). The assumption of the same measuring times, however, does not preclude the presence of incomplete data.

1.2 Literature review

This state of the art aims to explore the latest advances in spatio-temporal modelling approaches for the analysis of HTP data in the field of agriculture. We first review the literature in spatial models used in tra-

ditional agricultural experiments, where only a single measurement or observation is made at the end of the experiment. Such traits can be called endpoint traits. With the emergence of HTP technologies, agricultural experiments now include the temporal component, making the longitudinal evolution of genetic effects a key area of interest for plant breeders. Therefore, we examine the literature for longitudinal models in this context. However, to effectively model the genetic signal, we must also consider the temporal and spatial effects of environmental and design factors, which ultimately drives our search towards spatio-temporal modelling. We conclude our literature review indicating issues and challenges that guided the development of this thesis.

1.2.1 Spatial analysis in agricultural experiments

In agricultural experiments, it is well known that the phenotype of interest is spatially affected by micro-environmental factors. Experimental designs are used to counterbalance spatial heterogeneity in field trials (see, e.g., Brien et al., 2013; Hartung et al., 2019; Mead, 1997). In addition to experimental design, spatial models can be used to separate genetic and non-genetic effects properly. In this context, we assume that spatial models are defined in a two-dimensional row-column coordinate system. Spatial models for field trials are characterised by accounting for different sources of spatial variation (Gilmour et al., 1997): global (large-scale) variation across the field, local (small-scale) variation within the trial, and extraneous variation (related to the experimental procedure).

One alternative for spatial models is incorporating spatial variance-covariance structures in the errors (i.e., assuming spatially correlated noise). Examples of this kind of model are the separable autoregressive integrated model (ARIMA $\otimes$ ARIMA) by Cullis and Gleeson (1991). The problem with this model relies on the fact that differencing is used as an artefact to model the spatial trend, and as a consequence, a more complex variance-covariance spatial structure is imposed on the errors. Gilmour et al. (1997) and Verbyla et al. (1999) proposed a spatial mixed model to overcome this drawback. They use a separable autoregressive covariance structure (AR1$\otimes$AR1) to model local variation through the errors; add polynomial (fixed effects) or spline-based (fixed and random effects) functions of the spatial coordinates (i.e., rows and columns) to account for global variation; and include (when necessary) design factors effects (e.g., rows and columns) to model extraneous variation. In the bayesian framework, Besag and Higdon (1999) proposed a model with a spatially dependent structure in the prior; they specifically use a prior based on first differences along rows and columns to avoid nonstationarity (trend). The separable linear variance model (LV $\otimes$ LV) by Piepho and Williams (2010) follows the same logic: they propose to add a baseline row–column model (to capture extraneous variation) and a spatial structure through the errors.

The second alternative for spatial models is explicitly modelling the spatial variation and considering independent errors. In this direction we found proposals with smoothing methods. Green et al. (1985) proposed a least squares smoothing, but they approach the problem only in one-dimension. Durban et al.

(2003) use loess smoothers either as a sum of two one-dimensional trends (i.e., $f_1(\text{row}) + f_2(\text{col})$) or as a two-dimensional smooth surface (i.e., $f(\text{row}, \text{col})$). Finally, Rodríguez-Álvarez et al. (2018) propose a two-dimensional smooth surface by using anisotropic tensor product P-splines (Eilers & Marx, 2003) to explicitly model global and local spatial variation. They called their proposal SpATS (Spatial Analysis of Field Trials with Splines).

The separable autoregressive model (Cullis & Gleeson, 1991; Gilmour et al., 1997) has become the standard modelling strategy in field trials due to its popularity among applied breeders, mainly because of existing software (e.g. Genstat®, ASReml®, and ASReml®R, all of them paid software). However, this approach has some numerical problems (Piepho et al., 2015), making its application a cautious task that requires well-trained specialists. SpATS is the most recent approach for the analysis of large, complicated field trials using tensor product P-Splines; the proposal has proved to be very useful, and it is currently being used not only by researches and plant breeders but also by companies. Comparisons of these two models have shown that SpATS is a competitive and profitable approach (Andrade et al., 2020; Velazco et al., 2017), with the advantages that it can be easily adapted for HTP data, and the R-package `SpATS` is freely available for everyone.

1.2.2 Longitudinal analysis in agricultural experiments

In the context of HTP platforms, we are interested in modelling the temporal evolution of the genetic signal, where genotypes are represented by multiple plants in different replicates of the design. Therefore, we have a sample of plant curves as a function of time with a three-level hierarchical data structure (populations, genotypes nested in populations and plants nested in genotypes). Thus, we ask for hierarchical or multilevel longitudinal curve modelling approaches. Besides, traits are usually growth-related (e.g., canopy height and leaf area) but not necessarily (e.g., the efficiency of the photosystem), then flexible approaches that allow for non-linear relations are preferable.

Traditional analysis of growth-related curves uses parametric (non-linear) models. The logistic function is one of the most commonly used (Paine et al., 2012) to model individual curves. However, while growth processes theoretically follow a clearly defined pattern that may be modelled using a parametric function, the observed dynamics may deviate considerably due to, for example, temporal changes in environmental conditions (e.g. cold spells) or the application of treatments (e.g. irrigation events). More flexible models that overcome the limitations of parametric specifications for samples of curves have been proposed in the literature. Examples of data-driven methods include smoothing or penalised splines (P-splines; Eilers & Marx, 1996) and functional principal components analysis (FPCA; Ramsay & Silverman, 2005). It is worth noting that FPCA also contains penalised spline technology.

In the functional analysis framework, Greven and Scheipl (2017) provide an overview on functional re-

gression modelling and the software available. Di et al. (2009) extended the FPCA for data with a multilevel (or hierarchical) structure. They proposed the so-called MFPCA (multilevel functional principal component analysis). For the particular case of data from HTP platforms, plant growth dynamics analysed through MFPCA have been discussed in Xu, Li, and Nettleton (2018) and Xu et al. (2021). Xu, Qiu, et al. (2018) and Wang et al. (2020) use functional ANOVA, A. Montesinos-López et al. (2018) propose bayesian functional regression, and Miao et al. (2020) use FPCA, but none of them considers a hierarchical data structure of more than two levels (global mean and individual) in the analysis. For a specific analysis of HTP data, two R-packages result from these proposals: `implant` (not published; Wang et al., 2023) and `GFR` (not published; A. Montesinos-López et al., 2018). The packages `fda` (Ramsay et al., 2022) and `refund` (Goldsmith et al., 2022) are other more general options. In the latest, the `mfpca.sc()` is a specific function for MFPCA with penalised splines to smooth the covariance functions. A fast MFPCA (Cui et al., 2022) version is now available (jointly with the `mfpca.face()` function), which implements the fast covariance estimation method (FACE; L. Xiao et al., 2016).

Conversely, semiparametric regression models based on penalised splines and their connection with mixed models offer a rich framework for estimation and inference (Currie & Durban, 2002; Ruppert et al., 2003; Wand, 2003). An important reference here is the extension proposed by Brumback and Rice (1998), in which they used a sample of curves with nested and crossed effects. However, they considered fixed slopes and intercepts at the individual level, leading to a computational problem (identifiability). To overcome this drawback, Durban et al. (2005) proposed to use random coefficients, and Djeundje and Currie (2010) proposed to use a double penalty at the individual level (one for smoothness and one for identifiability). Particularly, Brien et al. (2020), Momen et al. (2019), O. A. Montesinos-López et al. (2017), and Moreira et al. (2020) are some of the works in the HTP context. The mixed model framework is also appealing for the variety of software available. Examples are the R-packages `nlme` (Pinheiro et al., 2019) and `lme4` (Bates et al., 2015) or the PROC MIXED procedure (SAS Institute Inc. 2015. SAS/STAT®, 2015). Also, the R-packages `mgcv` (Wood, 2017) and `gamm4` (Wood & Scheipl, 2020) can be used for that purpose.

1.2.3 Spatio-temporal analysis in agricultural experiments

In the two previous sections, we reviewed spatial and longitudinal models separately. In this section, we finally collect proposals that tackle the problem from both perspectives, temporal and spatial. The literature in this field is scarce. One possibility is to use stage-wise approaches. For instance, van Eeuwijk et al. (2019) propose first to estimate spatially adjusted genotypic means per time point and, in a second stage, they model the genotypic signal over time independently for each genotype. A similar approach is followed by Kar et al. (2020). Roth et al. (2021), describe a different stage-wise approach, where the temporal analysis is performed first, followed by the spatial correction. Stage-wise proposals have the advantage of being computationally manageable, but the problem relies on the loss of information between and within stages.

Therefore, it is of interest to develop approaches that allow modelling the spatial and temporal genetic and non-genetic variation in one stage. To the best of our knowledge, we only found one reference in this setting. Verbyla et al. (2021) have proposed modelling the genetic effects over time using factor analytic models and smoothing splines to model the non-genetic/residual effects over time and space. Nevertheless, the authors report their work as a "proof of concept" in the sense that fitting the models is very time-consuming. The authors use ASReml®R (Butler et al., 2018); nonetheless, they ask for scalable software.

1.3 Scope and challenges of the thesis

The scope of this thesis is in spatio-temporal models for HTP data to properly model the temporal evolution of the genetic effect on a specific phenotype while correcting for spatial and temporal environmental nuisance effects. We aim to combine statistical and computational methods that efficiently and adequately exploit the diversity and complexity of HTP data to

(a) Extract relevant information related to plant growth and development.

(b) Increase the biological understanding of plant systems.

(c) Help and support the decision-making process in plant breeding programs.

The literature review has revealed that spatio-temporal models for HTP data are a growing area of research as farmers and agronomists seek to optimise crop yields and improve farming practices through data-driven decision-making. Then the challenges of this research area include

(a) Determine the methodological path of our proposal and its implications (loss of information between and within stages and computational complexity). This involves deciding on the number of stages (i.e., whether to use a stage-wise approach or a single-stage approach) and the order of the stages (what component model first, the spatial or the temporal) for the approaches proposed.

(b) Select and combine appropriate statistical methods for the analysis of each component in the data (spatial and temporal).

(c) Take advantage of the spatio-temporal and hierarchical correlation data structure to avoid loss of information (when possible/necessary) between and within stages; improve the accuracy and robustness of the statistical models used; and better predict the data.

(d) Combine statistical and computational methods to deal with the size and complexity/dimensionality of the data. This requires efficient algorithms and software development that can handle the large datasets.

(e) Analyse real HTP data. HTP data collection and analysis itself poses several challenges: designing the experiment, collecting the data, image processing and cleaning the data can be time-consuming and resource-intensive and expensive. Although these challenges are out of the scope of this thesis, the investment in data acquisition is proof of the relevance of the projects in this field.

(f) Reproduce the dynamics of the real system as closely as possible using simulated data as an alternative/opportunity to assess the methods proposed in a set of different scenarios and configurations (Bustos-Korts et al., 2019).

(g) Develop scalable and user-friendly software for HTP data and make it accessible to the scientific community.

This thesis required a strong foundation in statistics, computational methods, and knowledge of plant systems. It involved developing and implementing spatio-temporal models, software design and development, and assessing and supporting with real and simulated data. The ultimate goal of this research is to improve plant breeding practices and contribute to global food security.

1.4 Thesis outline

The remainder of this thesis is organised as follows. We first introduce two motivating HTP data in Chapter 2. They are the greenhouse PhenoArch platform at INRAE Montpellier and the FIeld Phenotyping platform (FIP) at ETH Zürich. In this chapter, we use statistical data analysis to comment on different data characteristics such as the missingness pattern, the curve shape, the variability behaviour, possible factor interactions, and the spatial variation (and how it changes through time). For the FIP platform, three independent trials are available. Therefore, for this particular data, we additionally comment on the consistency of the genotypic performance across trials.

Chapter 3 contains the theoretical support of this thesis with concepts about P-splines and their extension to the multidimensional case by using tensor products. Two dimensions are used for spatial models (in the row and column directions) and three for spatio-temporal models (in the row, column and time directions). We elaborate on the connection between P-splines and mixed models as the estimation framework. This chapter outlines some essential technical details mainly used in Chapters 4 and 5.

The main methodological contributions of this thesis are presented in Chapters 4 and 5. To accomplish the aim of this thesis, we have proposed two approaches to the same problem. The first proposal divides the problem into two stages (two-stage approach). The first stage (spatial model) focuses on correcting the phenotypic data for experimental design factors and spatial variation, while the second stage (longitudinal and hierarchical model) aims to estimate the evolution over time of the genetic signal. The second proposal is to face the problem simultaneously (one-stage approach). That is, modelling the longitudinal evolution

of the genetic effect on a given phenotypic trait while accounting for the temporal and spatial effects of environmental and design factors (spatio-temporal and hierarchical model). We follow the same modelling philosophy throughout our work and propose P-splines-based models.

For our two-stage approach, we use the SpATS model separately for each measurement time in the first stage (similarly to van Eeuwijk et al., 2019). The phenotypic data are subsequently corrected by only considering the (estimated) sources of variation which are of interest plus the residual component (i.e., the measurement error). The purpose of this stage and subsequent correction is two-fold: (1) to remove nuisance spatial variation from the phenotypic data, and (2) to keep the data resolution for the second stage at the level of the experimental unit (through the incorporation in the correction of the residual component). This is one of the main differences to the proposals described by van Eeuwijk et al. (2019) and Kar et al. (2020), and it is routinely applied for data derived from the field phenotyping platform of ETH Zurich (Anderegg et al., 2020; Kronenberg et al., 2021; Perich et al., 2020). Since analyses are performed separately for each measurement time, our modelling strategy implicitly permits the spatial variation to differ among measurement times, i.e., it allows correcting for both the spatial and temporal evolution of environmental variables and experimental design factors. The second stage of our proposal focuses on modelling the genetic signal as a function of time for the corrected phenotype obtained in the first stage. Data for this stage consist of time-series of spatially corrected phenotypic trait measurements per experimental unit with a three-level hierarchical structure (populations, genotypes within populations, and plants within genotypes). We propose the use of P-spline hierarchical curve data models (psHDM) along the lines of the work by Durban et al. (2005) and Greven and Scheipl (2017).

Furthermore, we propose a one-stage approach to overcome the computational issues reported by Verbyla et al. (2021) and take advantage of all the available information given by the data structure. We use a spatio-temporal P-spline hierarchical curve data model (spatio-temporal psHDM). In particular, we generalise our two-stage modelling strategy to a full and one-stage spatio-temporal approach. We use the SpATS model as the base model and extend it to the spatio-temporal case, considering a three-level hierarchical data structure. Figure 1.2 depicts a comparative pictorial representation of both approaches in terms of inputs, modelling strategies and outputs.

For both approaches, we exploit the connection between P-splines and linear mixed models and propose to use computationally appealing tools that take advantage of the sparse structure of the matrices involved in the models to reduce computational complexity. As a result, we obtain estimated curves and their derivatives at the three levels of the hierarchy (populations, genotypes and plants). We use these curves to extract different time-independent characteristics. They can serve as inputs to subsequent statistical analyses aiming to model genotype-by-environment interactions in biologically endpoint traits, like yield, regarding underlying component traits (see, e.g., Moreira et al., 2020; van Eeuwijk et al., 2019).

Chapters 6 and 7 are devoted to comparing and assessing the performance of the one- and two-stage

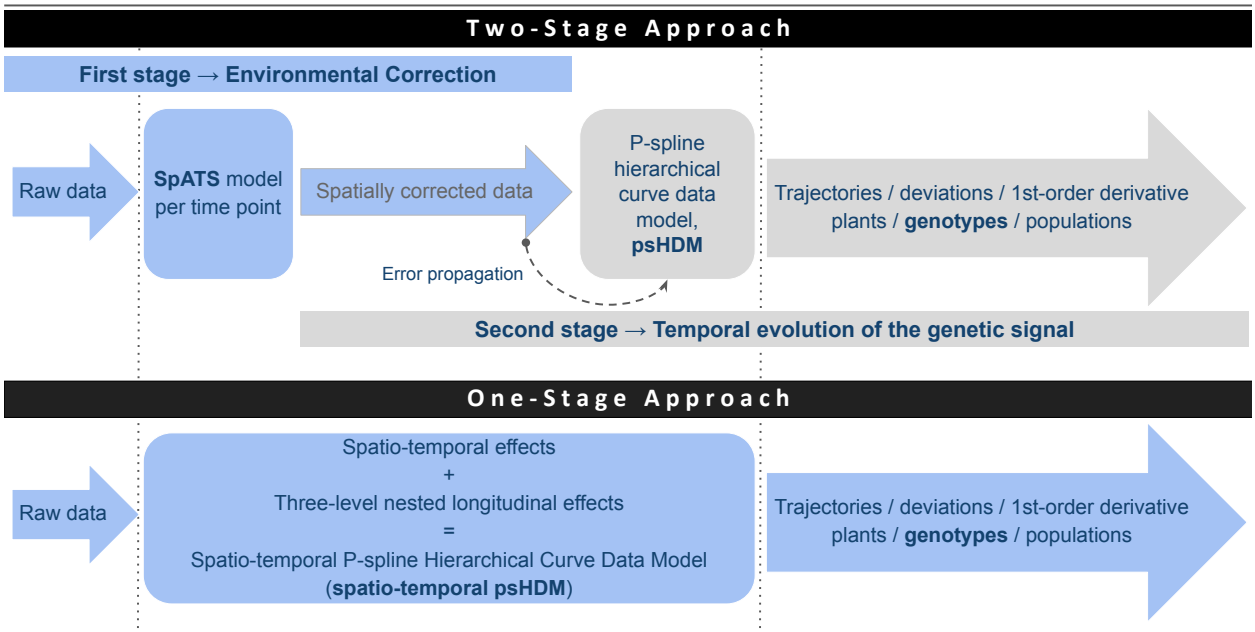


Figure 1.2: One and two-stage approaches in a pipeline.

approaches. In Chapter 6, spatio-temporal simulated data are used, while in Chapter 7, we analyse the two HTP data sets described in Chapter 2. To ensure consistency between the two approaches, we establish a common setting for both simulation and real-data analysis. The simulation study aims to support our research results and validate our approaches by exploring the impact of different variables and scenarios in a controlled environment. Simulation studies in this setting are relatively uncommon (see, e.g., Bustos-Korts et al., 2019; Roth et al., 2021), partly due to the (statistical and biological) complexity of the system, which makes the data generation mechanism a challenging task. We based the simulation setting and data generating model on our learnings with real HTP data analysis. To minimize bias, we propose a data generating model that is independent of the statistical models used for analysis. We are aware that our simulations involve some simplifications and assumptions that may not accurately reflect the complexities and specific features of real HTP data. Thus, to contrast the simulation findings, we analyse data from two different HTP platforms. Results with these two examples provide insights into factors not considered in the simulation, which will constitute important lines of future research.

As an effort to share our work and contribute to the progress of scientific knowledge, we have implemented the proposed approaches in the R language (R Core Team, 2023). For the one- and two-stage approaches, we propose functions to fit, predict, plot and extract time-independent characteristics. We show the functionalities of the code available in Chapter 8. By sharing our software code, we expect to ensure reproducibility of scientific results and to promote collaboration in the scientific community. We provide the user with one example to reproduce the results for the PhenoArch platform. Accordingly, it is our wish that other researchers can verify our findings, reproduce the analysis, and build upon our work. Moreover, we

aim to bring plant scientists closer to user-friendly statistical tools to support their decision-making process.

Most of the content of this thesis has been discussed in different academic spaces. Chapter 4 were published in Scientific Reports (Perez-Valencia et al., 2022). Chapters 5, 6 and 7 were comprised and submitted to the Journal of Agricultural, Biological and Environmental Statistics (JABES), but it is available in its BioXiv version (Perez-Valencia et al., 2023). All the functions presented in Chapter 8 are publicly available through the `statgenHTP` R-package (two-stage approach) and in https://gitlab.bcamath.org/dperez/htp_one_stage_approach (one-stage approach). Finally, this thesis ends with some conclusions in which the contributions of this thesis are summarised and some lines for future research are discussed in Chapter 9.

Chapter 2

Motivating examples: HTP data description

In this thesis, we analyse data of two experiments from two different HTP platforms: the PhenoArch platform (INRAE Montpellier; Cabrera-Bosquet et al., 2016) (greenhouse, Figure 2.1(a)) and three independent trials (2015, 2016 and 2017) performed at the FIP (FIeld Phenotyping) platform (ETH Zürich; Kronenberg et al., 2017) (field, Figure 2.1 (b)). In this chapter, we introduce the particularities of each platform and describe the data by means of a statistical descriptive analysis.

(a) PhenoArch platform at INRAE Montpellier (image source: INRAE).

(b) FIP platform at ETH Zürich (image source: ETH crop science).

Figure 2.1: Overview of two HTP platforms: **(a)** PhenoArch platform at INRAE Montpellier (greenhouse), and **(b)** FIP platform at ETH Zürich (field).

2.1 PhenoArch platform (INRAE Montpellier)

The PhenoArch platform is hosted at M3P, Montpellier Plant Phenotyping Platforms (https://www6.montpellier.inra.fr/lepse/M3P). It is composed of a conveyor belt structure of 28 lanes carrying 60 carts with one pot each (i.e. 1680 pots on a rectangular grid of 60 rows and 28 columns (see Figure 2.1(a)), plus a conveyor belt system that feeds the imaging or the watering units. Pots are daily moved to be imaged and/or watered. They are then moved back to the same positions and orientation, so that the plant position with respect to its neighbours is conserved throughout the experiment. The data analysed in this thesis correspond to an experiment including two different panels of commercial maize hybrids representative of breeding history in Europe and US during the last 60 years. This material covers a wide range of plant architecture, growth and development. A total of 90 genotypes were tested, 60 genotypes in Panel 1 and 30 genotypes in Panel 2. Each genotype was replicated between 4 (Panel 2) and 14 (Panel 1) times, approximately. All genotypes were tested under two levels of soil water content: mild water deficit (WD, soil water potential of -0.5 MPa) and retention capacity (WW, soil water potential of -0.05 MPa). The experiment was carried out in 2017 between April 13th and May 15th, corresponding to 103 and 135 days since January 1st (hereafter referred as DOY, Day of the Year), respectively. Red-green-blue (2056×2454) images taken from 13 views (12 side views from 30° rotational difference and one top view) were captured daily for each plant. Plant pixels from each image were segmented from those of the background and used for estimating the whole plant leaf area (among other features), as described in Brichet et al. (2017). Concerning the experimental design, a randomised complete block design was implemented for all four panel-by-water regime combinations, as depicted in Figure 2.2. This figure illustrates that the panel-by-water regime combinations were allocated by columns (blocks), and the assignment of genotypes to the experimental units (plants) within a block was done randomly. To be more specific, Panel 1 comprised ten complete blocks for WD and 14 for WW, while Panel 2 had four blocks for WD and four blocks for WW. In Panel 1, whole columns were utilised as blocks, whereas in Panel 2, half columns were employed as blocks. It is important to note that the blocks belonging to a particular panel-by-water regime combination were not all contiguous but were rather dispersed over the entire 60×28 grid to ensure comprehensive coverage of the platform's total variability as effectively as possible. The dataset consists of 32 leaf area measurements on 1656 plants ($1656 \times 32 = 52992$ observations, including missing data). We briefly characterise this dataset by a descriptive analysis:

1. **Time series curves with missing values at both the plant and genotype level** (i.e., with plants or even genotypes not measured for some times). Figure 2.3 shows the evolution over time of the raw leaf area data for plants in each panel by water regime. A total of 38930 (out of 52992) observations are available. There are between 5 to 13 missing values per plant curve (not all of them at the same time points), and 8 of the 90 genotypes are not measured at one time point. In addition, missing values are more present at the end of the experiment than at the beginning. That is, for the first 16 time points, we have 22234 (out of 26496) available data points, while for the second half of the

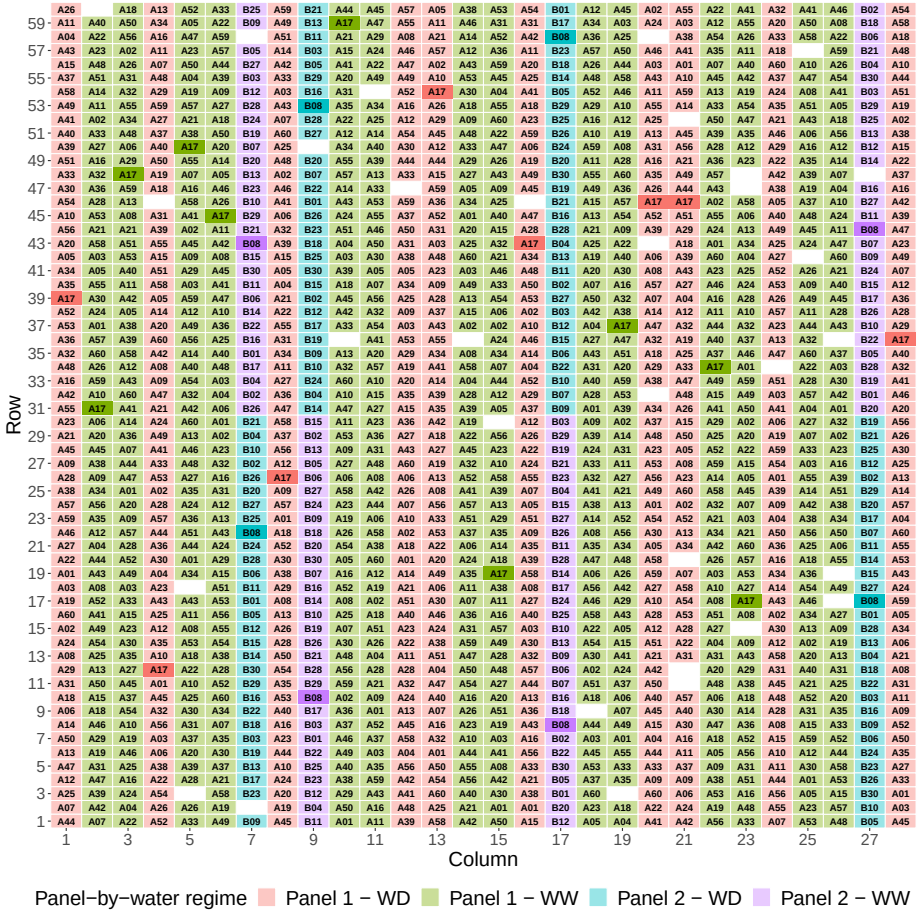


Figure 2.2: Descriptive analysis of the PhenoArch platform: Illustrative visualisation of the grid with the randomisation used for the PhenoArch experiment. The size of the grid is $R \times C = 60 \times 28$, for a total of $M = 1648$ plants (white spaces are missing values). Each cell represents a plant ($i = 1, \ldots, M$). Colours depict locations for plants in each panel by water regime. Highlighted colours depict replicates in two selected genotypes (as illustration): A17 and B08 in Panel 1 and 2, respectively, under the two water regimes.

observation period of the experiment, only 16696 values are accessible.

2. **The phenotypic trait (leaf area) is growth-related**. Plant-specific trajectories in Figure 2.3 show J-shape instead of S-shape curves, which means that information related with the stationary/steady phase is not available (e.g., asymptote or maximum trait as well as the time point at which this maximum is reached cannot be recovered).

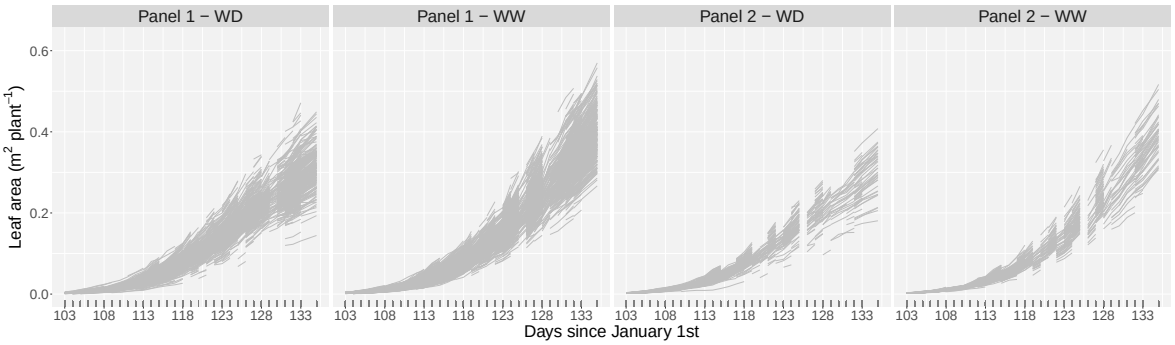


Figure 2.3: Descriptive analysis of the PhenoArch platform: Evolution over time of the raw leaf area.

3. **Between plants variability increases with time**. Figure 2.4 depicts the boxplots of the raw leaf area for each time point separately for plants in each panel (Panel 1 and 2) and water regime (WW and WD). In addition, this variability seems to be lower for plants in Panel 2 than those in Panel 1. With regard to the growth rate (slope, speed or average change of the leaf area through time), well watered (WW) plants seem to grow faster than those with water deficit (WD).

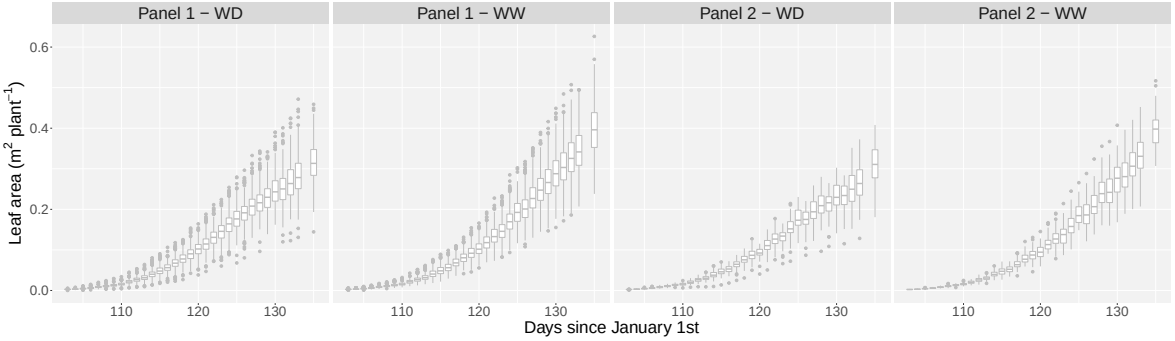


Figure 2.4: Descriptive analysis of the PhenoArch platform: boxplots of the raw leaf area grouped by time point.

4. **Different performance of genotypes because of the combination of genetic origin (panel) and water treatment.**. Figure 2.5 shows another view of Figure 2.4. In this case, we are more interested in showing the differences among panel by water regime combinations. Plants in the two panels behave similarly under each water regime, i.e., the leaf area of plants in Panel 1 - WD (red boxplots) / Panel 2 - WD (blue boxplots), and Panel 1 - WW (green boxplots) / Panel 2, WW (purple boxplots) are more similar. In addition, plants with the WW treatment (green and purple boxplots) seem to have

better performance (grow more) than those with the WD treatment (red and blue boxplots). These differences become greater through time (approximately from time point 127 onwards). Before that time point, differences in the leaf area seemed to be due to the panel and not to the water regime treatment.

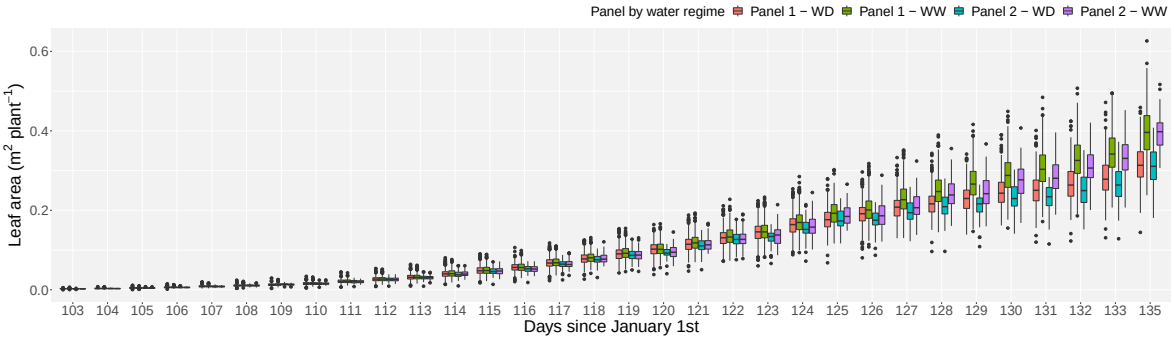


Figure 2.5: Descriptive analysis of the PhenoArch platform: Boxplots of the raw leaf area by panel and water regime treatment, grouped by time point.

Another perspective of this behaviour is shown in Figure 2.6. In this case, we emphasise differences by genotype (we use two genotypes, one per panel, as illustration). Genotype 43, in Panel 1, shows small differences in the leaf area between plants under the two water regime treatments, and these differences are more evident from time point 127 onwards. One important characteristic of the behaviour of this genotype is that plants in the WD (blue boxplots) treatment have a better (or almost equal) performance than those in the WW treatment (red boxplots), but this behaviour holds up to time point 126, after that period the effect is the opposite. The pattern for genotype 20 in Panel 2 is different: differences in the leaf area between plants under the two water regime treatments are larger, and they are evident from early time points; besides, plants in the WW treatment (red boxplots) always have a better performance than those in the WD treatment (blue boxplots).

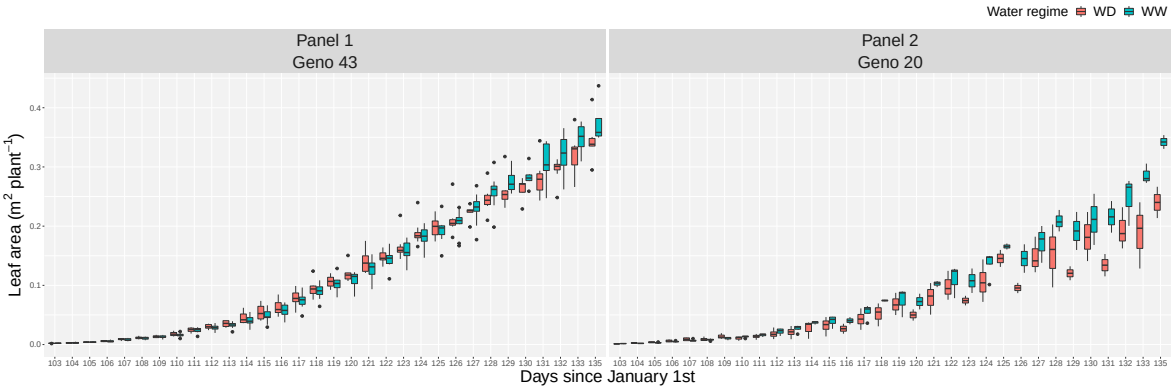


Figure 2.6: Descriptive analysis of the PhenoArch platform: Boxplots of the raw leaf area for two genotypes (one per panel, for illustration), grouped by time point and water regime treatment.

5. **Presence of nuisance spatial variation over time**. Figure 2.7 depicts the spatial distribution of the raw leaf area at four different measurements times (for illustration), where the spatial component smoothly changes over time.

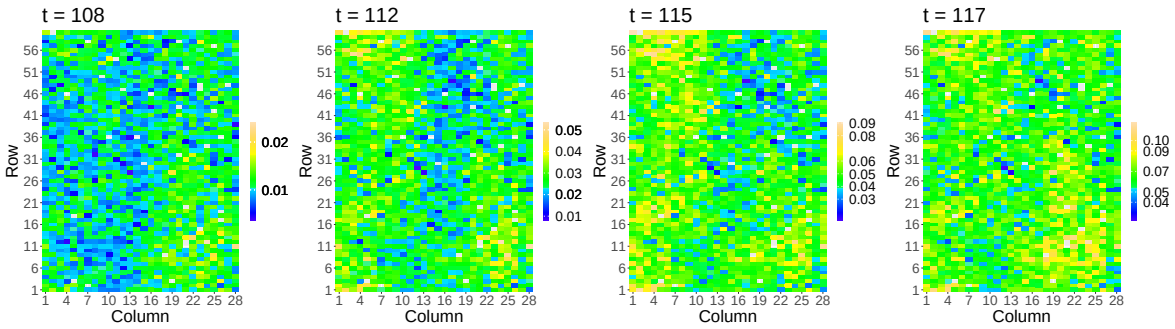


Figure 2.7: Descriptive analysis of the PhenoArch platform: Spatial distribution of the raw leaf area at four different measurements times (t = 108, 112, 115, 117 DOY) (white spaces are empty pots).

6. **Observations arising on the same plant, genotype or panel-by-water regime combination are serially correlated, and the covariance increases as a function of the shared grouping levels**. We follow the ideas in Di et al. (2009) to estimate the empirical covariance structure at the three levels of the hierarchy as shown in Figure 2.8: (i) the within plants covariance (Cov.plant), i.e., covariance within plant-specific trajectories; (ii) the within genotypes covariance (Cov.geno), i.e., covariance between plant-specific trajectories, calculated as the cross-sectional empirical means for all plants by genotype; and (iii) the within panel-by-water regime treatments (Cov.pop), i.e., covariance between genotypes of the same panel-by-water regime trajectories, calculated as the cross-sectional empirical means for all plants by panel-by-water regime. As expected, the covariance increases over time and it is higher for inner levels of the hierarchy (within plants covariance is the highest, followed by covariances within genotypes, and lastly covariances within populations).

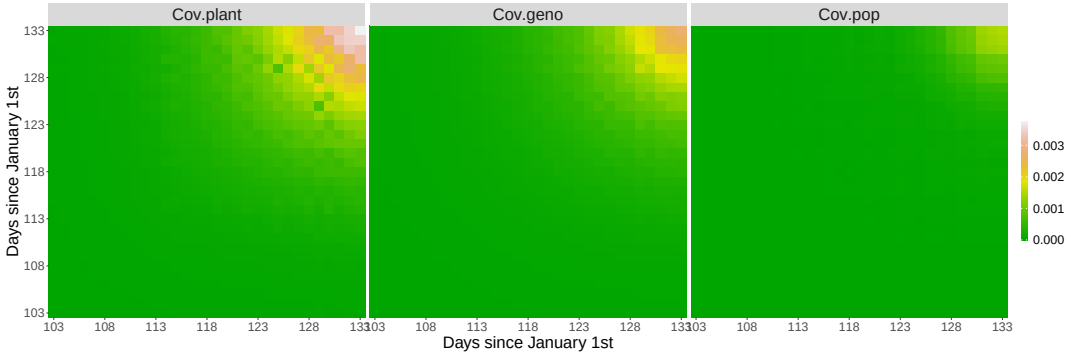


Figure 2.8: Descriptive analysis of the PhenoArch platform: Empirical covariance structure.

2.2 FIP platform (ETH Zürich)

The FIP platform, located at the ETH research station in Lindau-Eschikon (Switzerland), is a cable-suspended multi-sensor platform designed for automated, accurate and supervised high throughput data acquisition on an area of 1 hectare (Kirchgessner et al., 2016). From 2015 to 2017 (three different trials, one per year), the FIP platform was used to measure the development of canopy height on a diverse panel of European wheat genotypes (GABI wheat; Kollers et al., 2013), including a panel of Swiss varieties (Kronenberg et al., 2021; Kronenberg et al., 2017; Perich et al., 2020). Figure 2.1(b) shows the FIP platform with its crop rotation allocated to six different lots; for each trial, the wheat experiment was planted in two different lots, as shown in Table 2.1. Details on the experimental design can be found in Kronenberg et al. (2017) and Kronenberg et al. (2021). Figure 2.10 depicts the randomisation used for the three trials (2015, 2016 and 2017) of the FIP experiments. In short, the experimental unit was a plot to which the genotypes were allocated as the only treatment factor in an augmented 2D design. Checks (one for the 2015 and 2016 trials, and three for the 2017 trial) and test genotypes were allocated in a row-column design assuming that each row within a replicate (lot) received different environmental conditions (due to variability of crop husbandry measures, such as crop protection and fertilisation) while the upper, central and lower range of each lot received similar conditions (due to the similar slope direction within both lots). Canopy height measurements were carried out in irregular day intervals between February and July (depending on the trial, as indicated in Table 2.1), using a terrestrial laser scanner mounted on the FIP sensor head (Kronenberg et al., 2017). To analyse the experiment, we arranged the two replicates (lots) diagonally in a virtual grid with the number of rows and columns as indicated in Table 2.1. Additionally, the country in which a genotype was first inscribed into the European variety catalogue was also considered in the analyses. We used this information to allocate the genotypes to different wheat populations targeted at specific European regions. We will refer to these wheat populations (groups of genotypes) as regions of origin. Accordingly, plots are nested in genotypes, and genotypes are nested in regions of origin (wheat populations). The number of genotypes by region of origin and per trial is presented in Figure 2.9, where 313 genotypes were common to all three trials. The final dataset configuration of each trial for this platform is in Table 2.1. For the three trials 7 populations (regions of origin) are considered.

As for the PhenoArch platform data, we now describe some particularities of this dataset through a descriptive analysis:

1. **Time series curves at irregularly spaced time points**. Figure 2.11 shows the evolution over time of the raw canopy height by region of origin per trial. In contrast to the PhenoArch data, the FIP platform has 533 of missing observations, which is insignificant compared to the available data (42358 for the three trials). The characteristic of these time series curves relies on observations measured at irregular

Trial	2015	2016	2017
Lots	1 and 6	3 and 4	2 and 6
Row × Col (each lot)	17 × 21	18 × 20	18 × 21
Experimental dates (dd/mm)	27/04 - 25/06	11/04 - 04/07	27/02 - 30/06
Experimental dates (DOY)	117 - 176	102 - 186	58 - 181
Time points	17	21	23
Time between measurements (days)	3 - 5	2 - 8	2 - 13
Plots	680	678	720
Genotypes	322	319	334
Observations (including missing data)	11560	14238	16560

Table 2.1: Experimental configuration of each trial for the FIP platform (ETH Zürich)

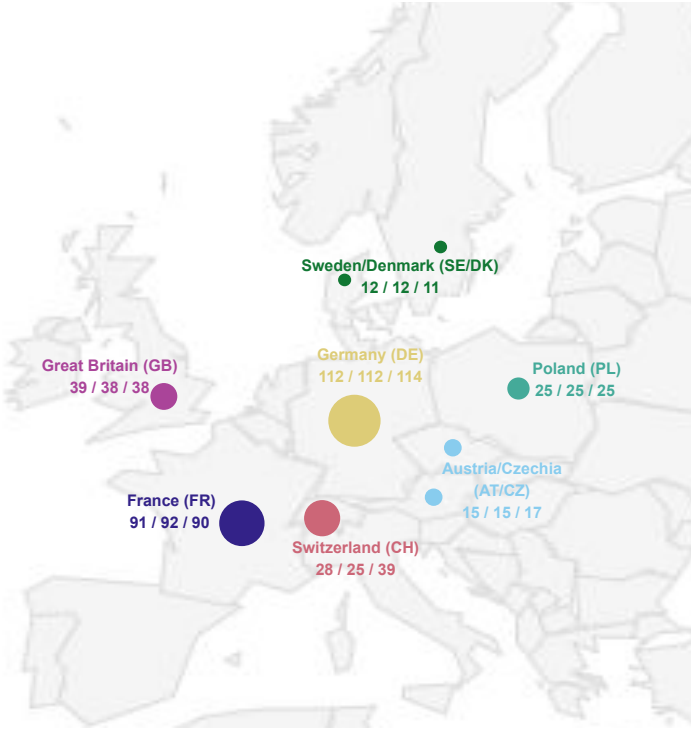


Figure 2.9: Number of genotypes by region of origin and for the three trials (2015, 2016 and 2017) for the FIP platform (ETH Zürich).

time points (Table 2.1 specifies the distance between time points for each trial, and $x-$axis in Figure 2.11 specifies time points at which observations were measured). Particularly, trait values for the 2017 trial are more spaced at the beginning and at the end of the period of observation. This characteristic also makes it difficult to calculate the empirical covariance structure as is depicted in Figure 2.8 for the PhenoArch dataset.

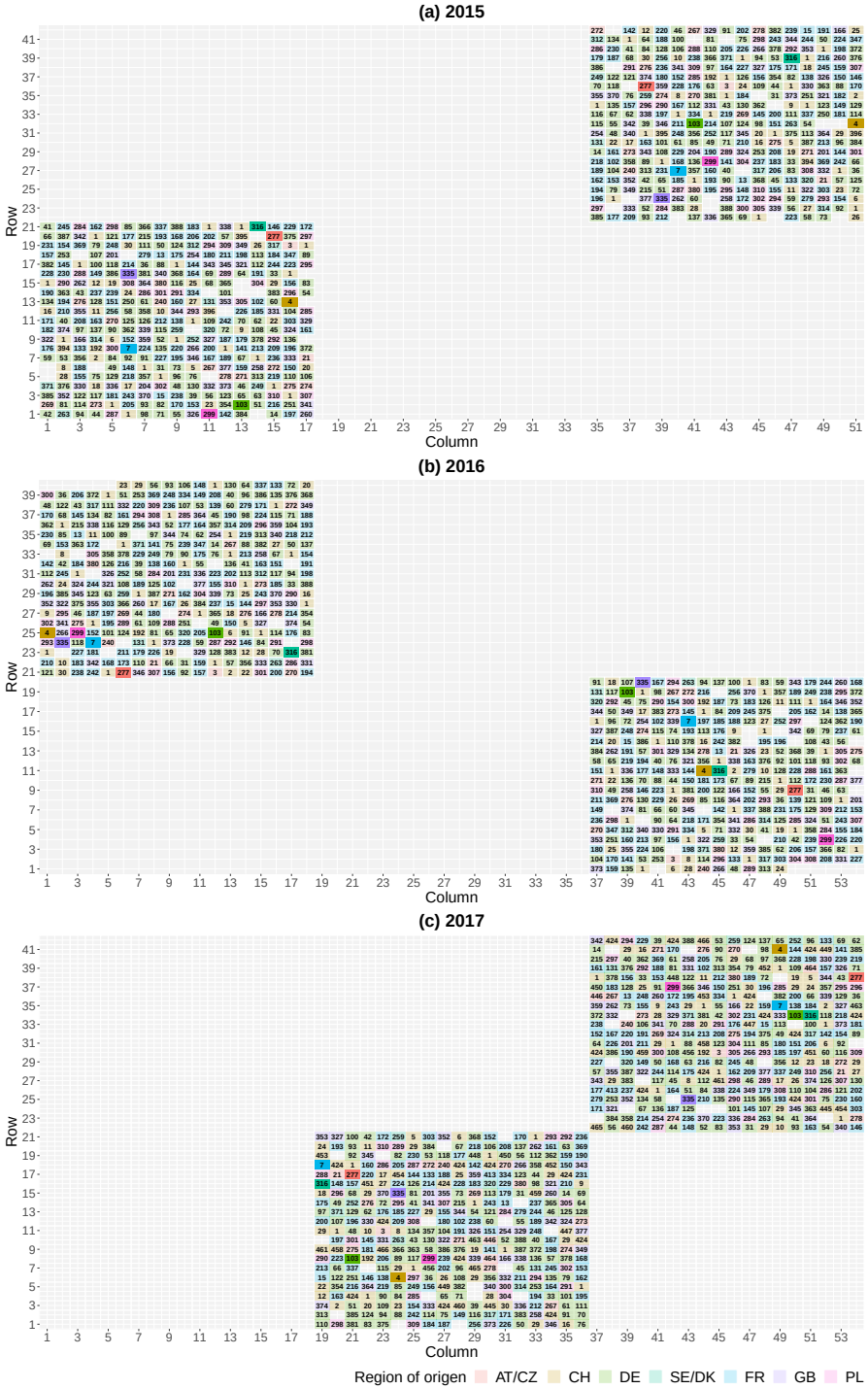


Figure 2.10: Descriptive analysis of the FIP platform: Illustrative visualisation of the grid with the randomisation used for the three trials of the FIP experiments, (a) 2015, (b) 2016 and (c) 2017. Each cell represents a plant ($i = 1, \ldots, M$), white spaces are missing values. Colours depict locations for plants in each region of origin. Highlighted colours depict replicates in one selected genotype by region of origin (as illustration).

2. **The phenotypic trait (canopy height) is growth-related**. Plant-specific trajectories in Figure 2.11 show S-shape curves. Nevertheless, information related to the initial/lag phase is unavailable for curves in trials 2015 and 2016.

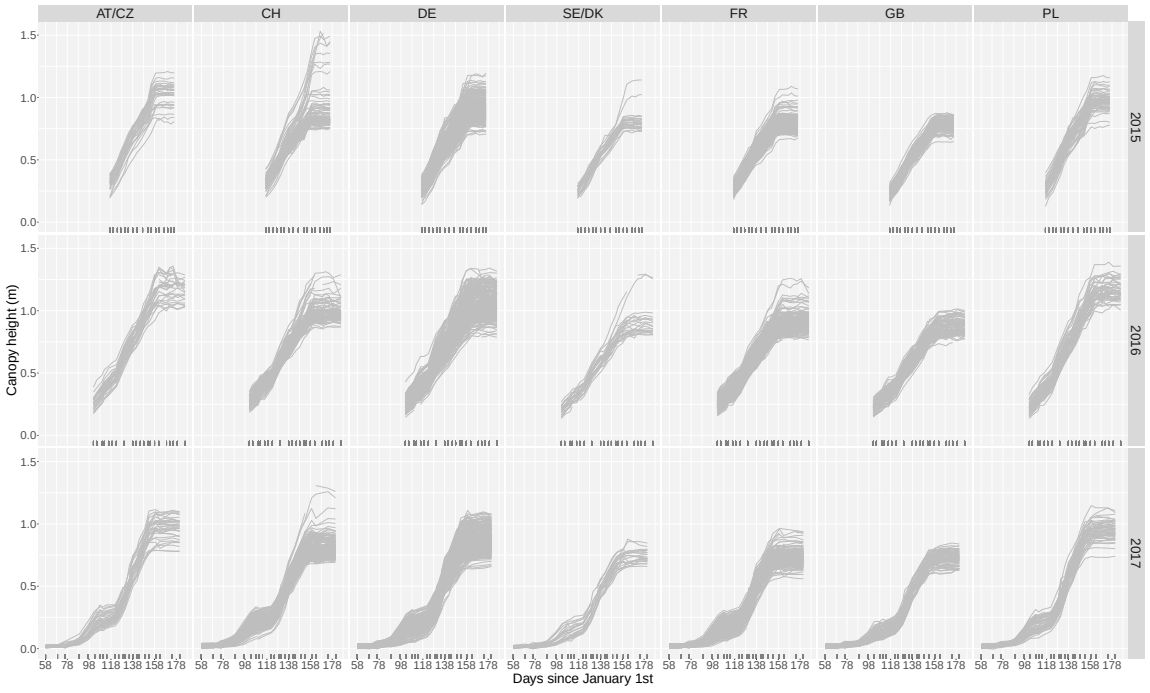


Figure 2.11: Descriptive analysis of the FIP platform: Evolution over time of the raw canopy height by region of origin (AT/CZ: Austria/Czechia, CH: Switzerland, DE: Germany, SE/DK: Sweden/Denmark, FR: France, GB: Great Britain, PL: Poland) for the three trials (2015, 2016 and 2017).

3. **Different performance of canopy height possibly due to regions of origin**. Figure 2.12 depicts the raw canopy height by trial, grouped by time point and region of origin. In this figure, we can observe the following:

   - **Between plants variability increases with time, and this variability differs among regions of origin**. For example, plant-specific trajectories from genotypes in regions of origin SE/DK and GB present lower variability than those in regions of origin AT/CZ and DE; this pattern remains more or less the same throughout the three trials.
   - Plants from genotypes in regions of origin PL and AT/CZ achieve (on average) larger canopy height values than those from SE/DK or GB; as for the between plants variability, this pattern remains more or less the same throughout the three trials. Nevertheless, the first maximum height time point seems different, even between trials. Generally, plants harvested in the 2016 trial achieved higher canopy height values.
   - **Clustering of genotypes by region of origin changes over time**. For example, for the three trials, three regional clusters are achieved in the steady/stable stage (up to time point 158, approx-

imately): GB-PL-SE/DK ("northern" Europe), are the clusters with the lowest canopy height values (as well as the lowest variability); CH-DE (central Europe) is the cluster with middle canopy height values; and finally, AT/CZ-PL (eastern Europe) is the cluster with the highest canopy height values. This grouping is consistent with the geographic location of the regions of origin in Europe (see map in Figure 2.9). Additionally, observe that clustering seems to change through time according to the growth curve stages (i.e., clusters are more or less stable in the lag phase, exponential phase, and steady phase of each trial).

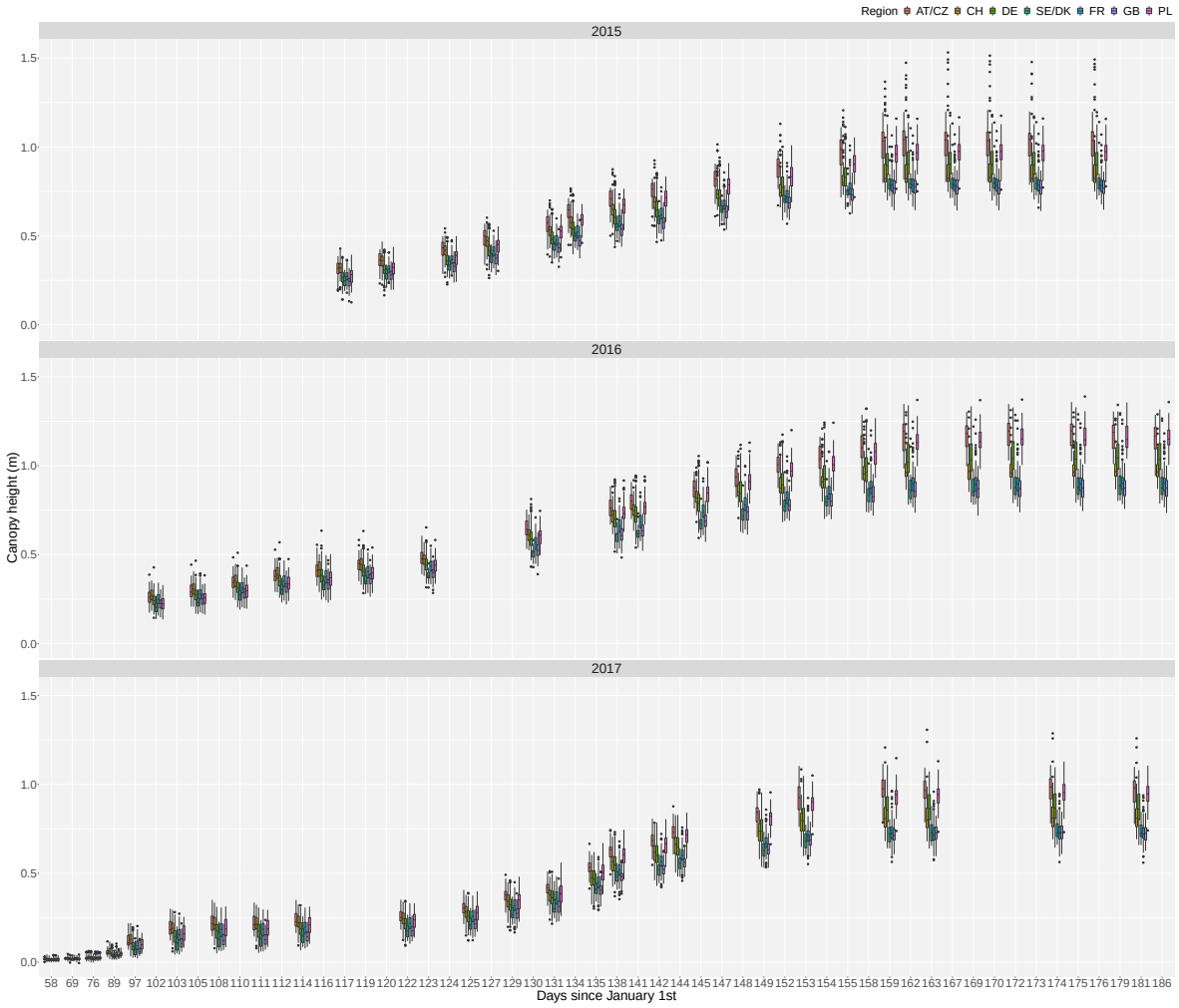


Figure 2.12: Descriptive analysis of the FIP platform: Boxplots of the raw canopy height by trial, grouped by time point and region of origin.

4. **Genotype by trial interaction** (this can be done for the 313 genotypes common to all three trials). In Figure 2.13, we compare the raw canopy height over time for the two replicates of one genotype by region of origin across trials (for illustration). For instance, for the 2016 trial (green boxplots), genotype 278 has a better average performance than genotype 110, but this behaviour is not the same for the other two trials. Additionally, this performance can change over time, e.g., for the 2017 trial

(blue boxplots) the average performance of genotype 278 is better than the one for genotype 110 during the lag and exponential phases, but the behaviour is opposite in the stable phase. In this case, we compare just two genotypes, but even when common genotypes were used, information about "best" (or "worst") genotypes that are common to the three trials, or genotype improvement across trials could be extracted. We also can observe differences in the plant-to-plant variation (or within genotype variation) through trials. For instance, genotypes 8 and 166 have higher within genotype variation for the 2016 trial than for the 2015 and 2017 trials.

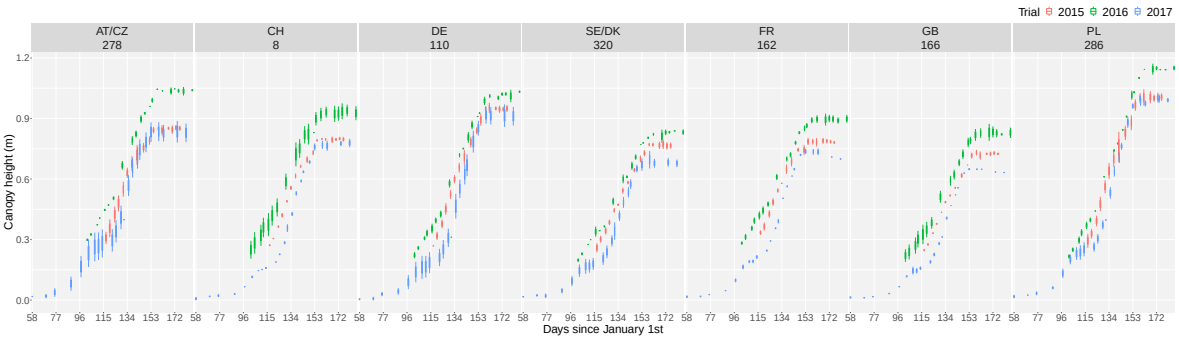


Figure 2.13: Descriptive analysis of the FIP platform: Evolution over time of the raw canopy height for the two replicates of one genotype by region of origin (as illustration) for the three trials (2015 in red, 2016 in green, and 2017 in blue).

5. **Effect of external environmental factors**. For example, Figure 2.14 depicts the raw plant-specific trajectories (grey lines) and the mean temperature (blue line) for the 2017 trial. The temperature has a particular effect during the cold period in April (DOY 110-120) on the "standard" S-shape of the plant-specific curves (that is, they are temporarily disrupted).

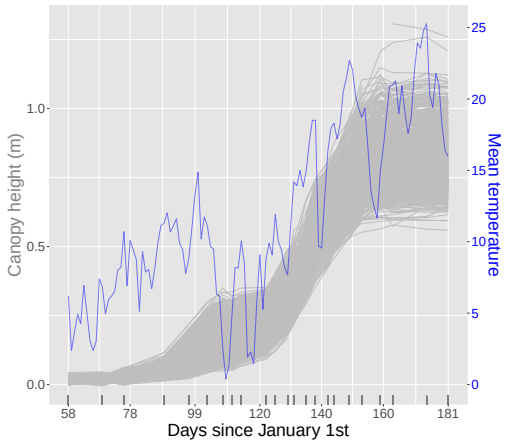
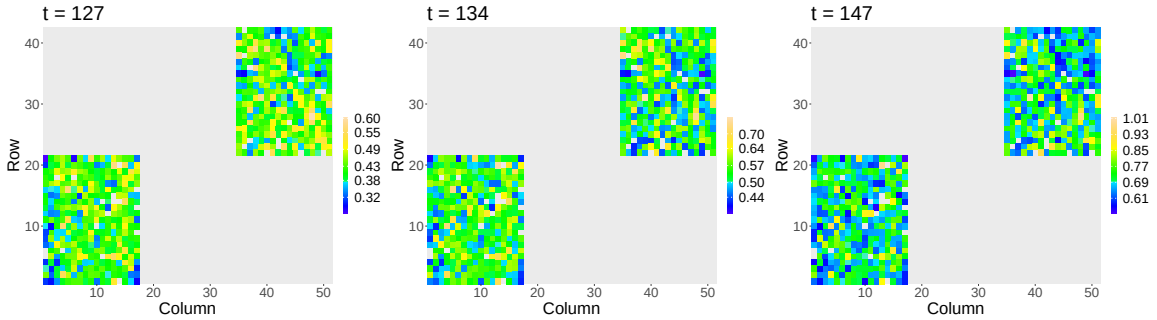


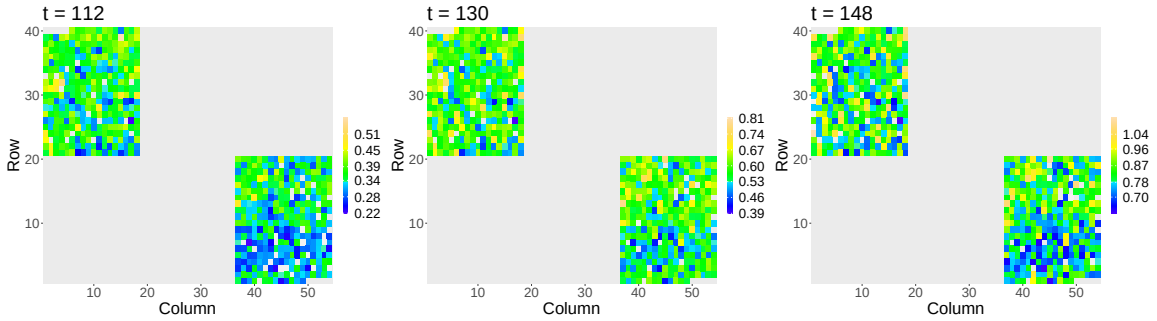
Figure 2.14: Descriptive analysis of the FIP platform: raw plant-specific trajectories (grey lines) vs. mean temperature (blue line) for the 2017 trial.

6. **Presence of nuisance spatial variation over time**. Figure 2.15 depicts, for the three trials, the spatial distribution of the raw canopy height at three measurement times (for illustration). As for the

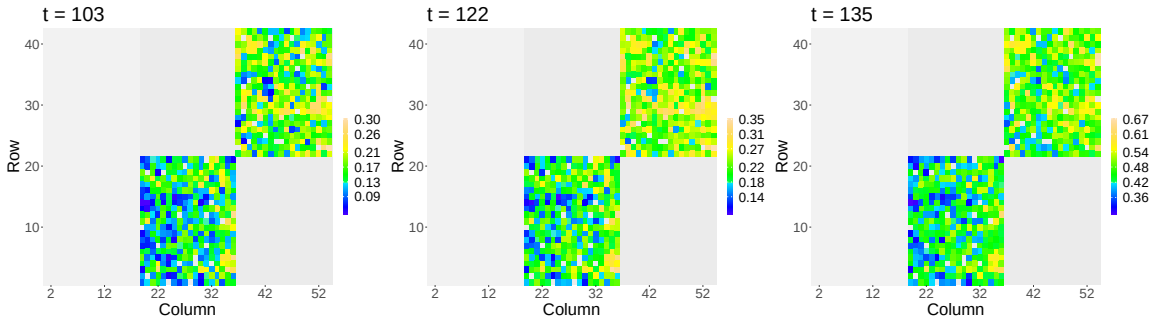
PhenoArch dataset, the spatial component smoothly changes over time. In this case, the additional lot effect (experimental design factor) can be observed in these spatial plots. For instance, for the 2017 trial in Figure 2.15(c), replicates in lot 2 (see Figure 2.1 (b) to better identify the position of the lots in the field) have smaller canopy height values than those in lot 6. We note that, an additional characteristic of this experiment is the non-adjacent lots.



(a) 2015



(b) 2016



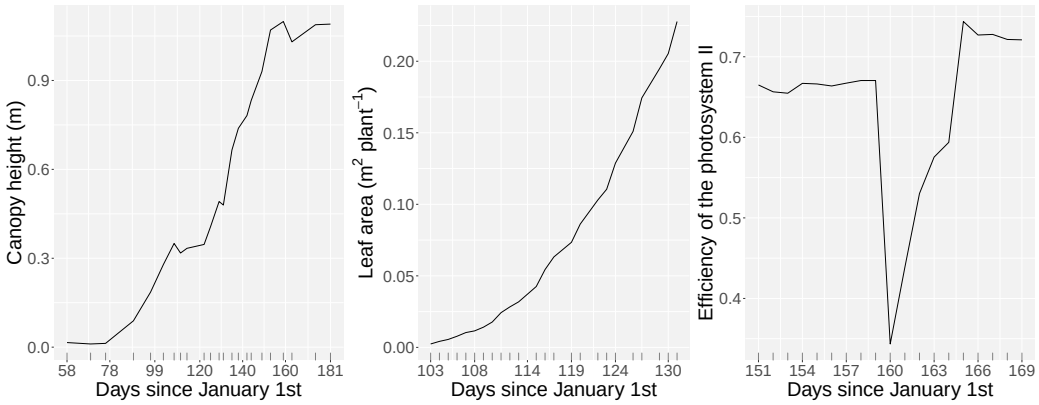
(c) 2017

Figure 2.15: Descriptive analysis of the FIP platform: Spatial distribution of the raw canopy height at three measurement times for all three trials: **(a)** 2015 ($t = 127, 134, 147$ DOY), **(b)** 2016 ($t = 112, 130, 148$ DOY), and **(c)** 2017 ($t = 103, 122, 135$ DOY). We note that columns 1 to 19 are missing for the 2017 trial (lots 2 and 6 are sown as indicated in Figure 2.1(b)).

Chapter 3

P-splines, tensor products and mixed models

In the context of HTP data, we are interested in modelling a phenotype of interest (e.g., plant height, canopy cover, leaf area index, ear and tiller counts, canopy temperature), as a function of some covariates (e.g., populations of genotypes, genotypes, time, row and column position of the plant/plot in the field/greenhouse). For simplicity, let's consider the one-dimensional case in which we want to study, e.g., the evolution over time of a phenotype of interest for one plant of a given genotype. As illustration, Figure 3.1 depicts the temporal evolution of three different phenotypes from three different experiments.



(a) Canopy height (b) Leaf area (c) Efficiency of the photosystem II

Figure 3.1: Temporal evolution of three different phenotypes for one plant on a given genotype (as illustration) for three different experiments: **(a)** Canopy height (FIP platform, ETH, Zürich, Switzerland), **(b)** Leaf area (PhenoArch platform, Montpellier, France), and **(c)** Efficiency of the photosystem II (PhenoVator platform; data available in the R-package `statgenHTP`, Millet et al., 2022).

To model the relationship between the phenotype and time we consider the following regression model

$$y(t) = f_{\mathrm{T}}(t) + \varepsilon(t),\ \varepsilon(t) \sim N(0, \sigma^2),\ t \in \{t_1, \ldots, t_n\}, \tag{3.1}$$

where $f_{\mathrm{T}}(t)$ is an unknown function, and $\varepsilon(t)$ is an error with variance σ^2. We can be tempted to assume a parametric form for $f_{\mathrm{T}}(t)$ as first approach. For instance, to model the temporal evolution of the canopy height in Figure 3.1(a), we can start describing a linear relation. Clearly, it is not linear, then we can propose a polynomial function. Another alternative is to use a logistic function, which has demonstrated to be a good approach for (S-shaped) growth curves (Paine et al., 2012). Nevertheless, in the context of HTP data, parametric approaches can be very restrictive (and particular) to the shape of the phenotypic curve (e.g., Figure 3.1 shows three different shapes). Instead, more flexible approaches can be applied that overcome the limitations of parametric specifications. One example are the well-known penalised splines (P-splines, Eilers & Marx, 1996). Figure 3.2 shows estimated curves for the data shown in Figure 3.1(a) based on different approaches. This thesis is focused on P-splines. It is a very appealing approach that thanks to its connection with the linear mixed model, offers a rich framework for estimation and inference (Currie & Durban, 2002; Wand, 2003), even in the presence of missing data. Moreover, we can easily obtain the derivatives of the estimated curves, which can provide important insights for the decision-making process in plant breeding.

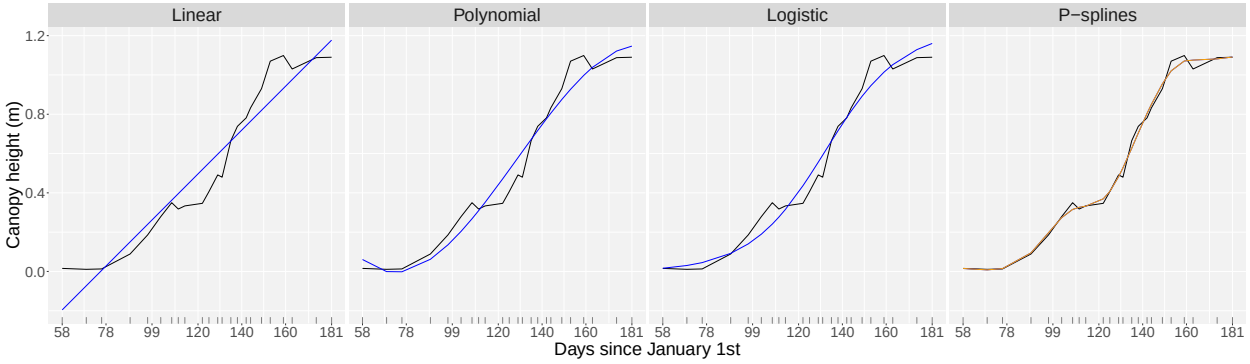


Figure 3.2: (a) Estimated curves for the temporal evolution of the canopy height shown in Figure 3.1(a). Four different approaches are depicted: linear regression, polynomial regression, logistic function and P-splines. Raw canopy height is in black and estimated curves are in blue. The estimated curve using P-splines is in orange to indicate our choice.

The remainder of this chapter is devoted to a detailed description of the specification of P-spline-based models in one, two and three dimensions. These models form the basis of the proposals presented in Chapters 4 and 5, that constitute the methodological contribution of this thesis. In this Chapter we also discuss and present the procedures for the estimation of P-spline models. This includes their mixed model formulation, parameters and derivatives estimation, and calculation of pointwise confidence intervals.

3.1 P-splines overview

P-splines were introduced by Eilers and Marx (1996), but more details about this smoother can be found in their recent book (Eilers & Marx, 2021). The two principal elements of P-splines are B-splines (for more details, see de Boor, 1978) and discrete penalties on the regression coefficients. We first describe P-splines for one-dimension and later present the extension to multidimensional P-splines. Particularly, due to the spatial and temporal dimensions that characterise the kind of HTP data we deal with, we will naturally focus on spatio-temporal models. For P-splines in one dimension, we will specify a temporal model. Then, for the bi-dimensional case, we will explicitly comment on spatial models, and for three dimensions, we will discuss on spatio-temporal models.

3.1.1 P-splines in one dimension: temporal effect

Following the problem presented above, we are interested in studying the temporal evolution of a phenotype of interest $y(t)$, for one plant on a given genotype. Let $f_{\mathrm{T}}(t)$ in (3.1) be approximated by a linear combination of b_1 known B-splines basis functions of degree q defined over a sequence of equally-spaced knots, i.e.,

$$f_{\mathrm{T}}(t) = \sum_{k_1=1}^{b_1} B_{k_1}(t; q)\theta_{k_1}, \tag{3.2}$$

where $\boldsymbol{\theta}_{\mathrm{T}} = (\theta_1, \ldots, \theta_{b_1})^T$ is a vector of unknown regression coefficients that controls the shape of the curve. In matrix form we write,

$$\boldsymbol{f}_{\mathrm{T}} = \boldsymbol{B}_{\mathrm{T}}\boldsymbol{\theta}_{\mathrm{T}},$$

where $\boldsymbol{f}_{\mathrm{T}} = (f_{\mathrm{T}}(t_1), \ldots, f_{\mathrm{T}}(t_n))^T$, and $\boldsymbol{B}_{\mathrm{T}}$ is a (temporal) B-spline design matrix with n rows (i.e., the number of observations/time points for one plant) and b_1 columns, i.e.,

$$\boldsymbol{B}_{\mathrm{T}} = \begin{pmatrix} B_1(t_1; q) & B_2(t_1; q) & \cdots & B_{b_1}(t_1; q) \\ \vdots & \vdots & \ddots & \vdots \\ B_1(t_n; q) & B_2(t_n; q) & \cdots & B_{b_1}(t_n; q) \end{pmatrix}. \tag{3.3}$$

Note that $(\boldsymbol{B}_{\mathrm{T}})_{jk_1} = B_{k_1}(t_j; q)$ is the k_1th B-spline basis function evaluated at time point t_j, and q is the degree of the B-spline basis. A graphical representation of B-spline basis functions for different values of q is shown in Figure 3.3.

Under specification (3.2), model (3.1) is purely parametric, and $\boldsymbol{\theta}_{\mathrm{T}}$ is estimated by minimising the residual sum of squares $\mathrm{RSS}_{\mathrm{T}} = \|\boldsymbol{y} - \boldsymbol{B}_{\mathrm{T}}\boldsymbol{\theta}_{\mathrm{T}}\|^2$. Then the (unpenalised) estimator of $\boldsymbol{\theta}_{\mathrm{T}}$ accepts a closed-form expression with the explicit solution $\hat{\boldsymbol{\theta}}_{\mathrm{T}} = (\boldsymbol{B}_{\mathrm{T}}^T\boldsymbol{B}_{\mathrm{T}})^{-1}\boldsymbol{B}_{\mathrm{T}}^T\boldsymbol{y}$. When working with B-splines, two choices must be made: (i) the B-spline degree, q, and (ii) the number of basis functions, b_1. In Figure 3.4(a) estimated curves using B-splines of degree 1 to 4 are compared. Given the nature of our data and that we are interested

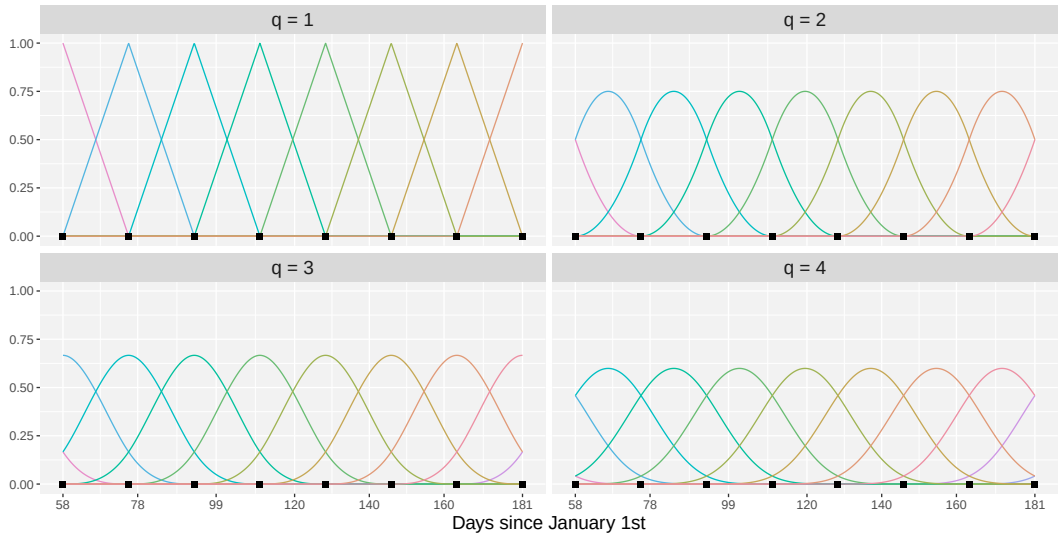


Figure 3.3: Illustrative visualisation of B-splines basis functions with 7 segments, equally-spaced knots (black points along time), and different degree-orders ($q = 1, 2, 3, 4$).

in estimating derivatives, along this thesis we will work with cubic B-splines (i.e., $q = 3$). Figure 3.4(b) depicts results for estimated curves using cubic B-splines for different basis dimensions, b_1. In general, the higher this number, the wigglier the estimated curve.

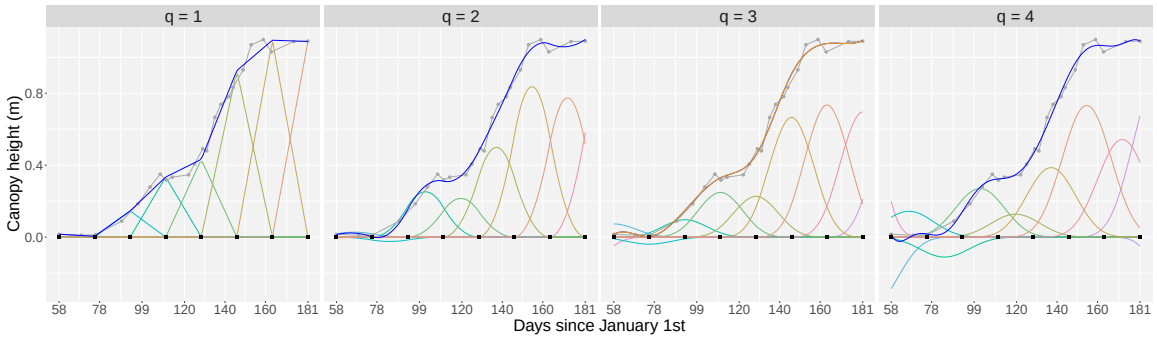
One option to control the amount of flexibility given by the choice of a large number of basis functions, b_1, while keeping a smooth fit, is to penalise the regression coefficients, $\boldsymbol{\theta}_{\mathrm{T}}$. In P-splines, the idea is to form (the sum of squares of) differences of order d on the vector of regression coefficients, i.e.,

$$\sum_{k_1=d+1}^{b_1} \left(\Delta^d \theta_{k_1}\right)^2 = \boldsymbol{\theta}_{\mathrm{T}} \boldsymbol{D}_1^T \boldsymbol{D}_1 \boldsymbol{\theta}_{\mathrm{T}}. \tag{3.4}$$

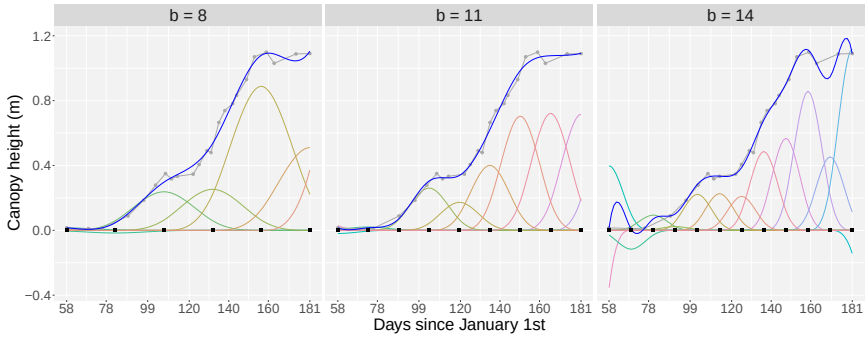
Here, Δ^d forms differences of order d, i.e., $\Delta\theta_{k_1} = \theta_{k_1} - \theta_{k_1-1}$, $\Delta^2\theta_{k_1} = \theta_{k_1} - 2\theta_{k_1-1} + \theta_{k_1-2}$, and so on for higher d; and $\boldsymbol{D}_1$ is simply the matrix representation of Δ^d. Note that if neighbouring elements of $\boldsymbol{\theta}_{\mathrm{T}}$ are similar/dissimilar, then the coefficient differences will be small/large. As illustration, Figure 3.5 shows second order differences (i.e., $d = 2$) on adjacent coefficients. Thus, (3.4) can be used to measure how "rough" $\boldsymbol{\theta}_{\mathrm{T}}$ is and to penalised its estimate. To that end, the residual sum of squares $\mathrm{RSS}_{\mathrm{T}}$ is modified, and a penalised residual sum of squares is considered instead

$$\mathrm{RSS}_{\mathrm{T}} + \underbrace{\lambda_1 \boldsymbol{\theta}_{\mathrm{T}}^T \boldsymbol{P}_{\mathrm{T}} \boldsymbol{\theta}_{\mathrm{T}}}_{\text{Penalty term}}, \tag{3.5}$$

where $\boldsymbol{P}_{\mathrm{T}} = \boldsymbol{D}_1^T \boldsymbol{D}_1$, and λ_1 is a smoothing parameter that sets the weight of the penalty: the larger λ_1, the smoother the result will be. In other words, λ_1 controls the trade-off between fidelity to the data (when λ_1 is small) and smoothness of the function estimate (when λ_1 is large). It is easy to show that, for a given λ_1,



(a) B-splines-based estimated curves with different order, q.



(b) Cubic B-splines estimated curves with different basis dimensions, b_1.

Figure 3.4: B-splines-based estimated curves (in blue), for the canopy height shown in Figure 3.1(a), with equally-spaced knots (black points along time), together with the B-splines basis functions, scaled by the estimated coefficients, $\hat{\boldsymbol{\theta}}_{\mathrm{T}}$, and **(a)** different degree orders ($q = 1, 2, 3, 4$) with 7 segments ($nseg = b_1 - q$); estimated curves for $q = 3$ are in orange to indicate our choice, **(b)** different number of basis functions ($b_1 = 8, 11, 14$), and $q = 3$.

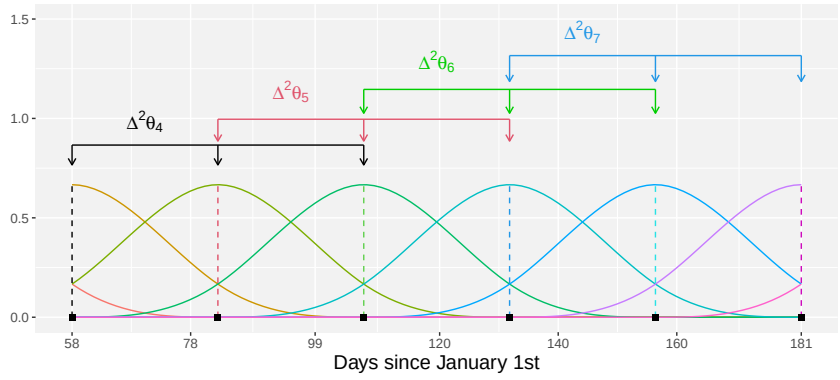
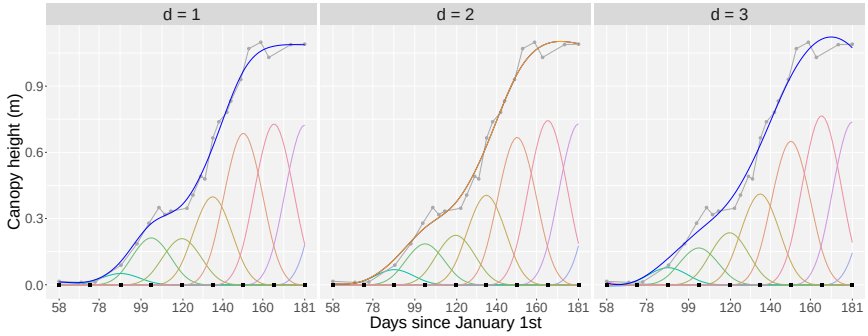


Figure 3.5: Illustrative visualisation of second order differences on adjacent coefficients of $b_1 = 8$ cubic B-splines basis functions.

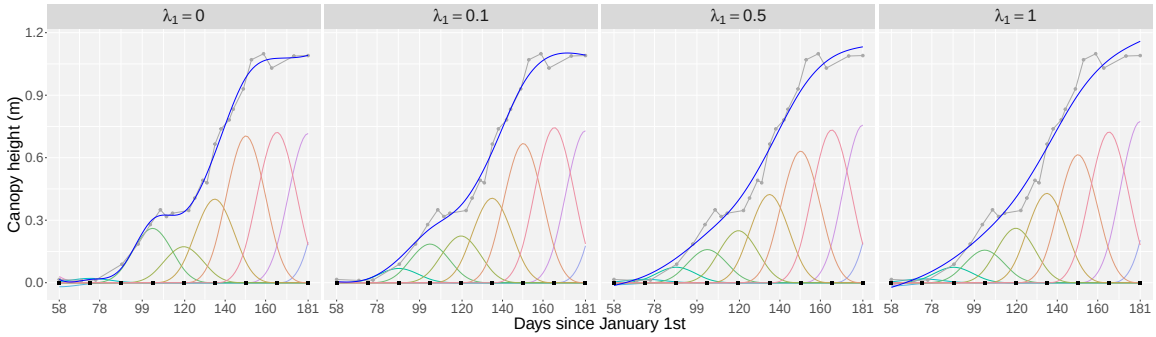
the solution of (3.5) also accepts a closed-form expression

$$\hat{\boldsymbol{\theta}}_{\mathrm{T}} = (\boldsymbol{B}_{\mathrm{T}}^{T}\boldsymbol{B}_{\mathrm{T}} + \lambda_1 \boldsymbol{P}_{\mathrm{T}})^{-1}\boldsymbol{B}_{\mathrm{T}}^{T}\boldsymbol{y}. \tag{3.6}$$

Figure 3.6(a) depicts estimated curves using P-splines with different d values; in this thesis, we will use second-order differences (i.e., $d = 2$). Figure 3.6(b) shows results using different values for the smoothing parameter; the larger the λ_1, the smoother the result. After the number of B-spline basis functions is fixed, the only tuning mechanism for smoothness is the strength of the penalty, i.e., the value of the smoothing parameter λ_1. Accordingly, a critical issue is setting the right value for λ_1, which we like to see determined by the data. We discuss this point in detail in Section 3.2.



(a) P-splines-based estimated curves with different *d*order difference.



(b) P-splines-based estimated curves with different smoothing parameter λ_1.

Figure 3.6: P-splines-based estimated curves (in blue), for the canopy height shown in Figure 3.1(a), with equally-spaced knots, along with the cubic B-splines basis functions, scaled by the estimated (penalised) coefficients, $\hat{\boldsymbol{\theta}}_{\mathrm{T}}$, and (**a**) different d-order differences ($d = 1, 2, 3$), $b_1 = 11$, $q = 3$ and $\lambda_1 = 0.1$; estimated curves for $d = 2$ are in orange to indicate our choice **(b)** different values for the smoothing parameter ($\lambda_1 = 0, 0.1, 0.5, 1$), $b_1 = 11$, $q = 3$, $d = 2$.

3.1.1.1 Derivatives

When working with B-splines, derivatives from curves of the form (3.2) can be easily obtained using the formula given by de Boor (1978). In general, derivatives of order h can be computed as follows (Eilers & Marx, 2021)

$$\frac{\partial^h}{\partial t^h} f_{\mathrm{T}}(t) = \frac{\partial^h}{\partial t^h} \sum_{k_1=1}^{b_1} B_{k_1}(t; q)\theta_{k_1} = \sum_{k_1=h+1}^{b_1} B_{k_1}(t; q-h)(\Delta^h \theta_{k_1})/\omega^h, \tag{3.7}$$

where $B_{k_1}(t; q-h)$ is a B-spline basis function of degree $q-h$ (it can be obtained explicitly), Δ^h forms differences of order h (in a similar fashion to expression (3.4)), and ω is the length of the domain of $\boldsymbol{t}$ divided by the number of segments. Note that the condition $h \leq q$ must be satisfied, otherwise the derivative will be zero. In our case, as we are explicitly working with cubic B-spline basis (i.e., $q = 3$), we can compute derivatives up to order $h = 3$. Note that based on equation (3.7) it seems reasonable to estimate the hth order derivative of function f_{T} by plugging into (3.7) an estimate of $\hat{\boldsymbol{\theta}}_{\mathrm{T}}$ using (3.6).

3.1.2 Multidimensional P-splines and tensor products

We now move to multidimensional P-splines where more than one effect will be considered to model a phenotype of interest. For instance, in two-dimensions (rows and columns) we will talk about the spatial effect, and in three-dimensions (rows, columns and time) we will comment on spatio-temporal effects. As a first approach linear additive functions (i.e., sum of one-dimensional functions) without interactions could be used. Instead, to allow for interactions between the covariates, we will use tensor products of one-dimensional cubic B-spline basis while simultaneously penalising the coefficients on each dimension considered (Eilers & Marx, 2021). Nevertheless, one important computational implication of the use of tensor products is that the size of the problem grows with the number of dimensions required.

3.1.2.1 Two-dimensional P-splines: Spatial effect

As we said before, our motivation to use two-dimensional P-splines comes from the need to study the spatial pattern of a phenotype of interest y_i ($i = 1, \ldots, M$) given by the location of M plants, in a rectangular grid (two-dimensional space) with R rows and C columns. We use r and c for indices of row and column position, respectively ($r = 1, \ldots, R$; $c = 1, \ldots, C$). With a slight abuse of notation, we use $r(i)$ and $c(i)$ to denote the row and column position of the ith plant, respectively. As example, Figure 2.15 depicts the spatial distribution of the raw canopy height for the FIP platform data, at one measurement time for all three trials. As a first approach, an interaction model in two dimensions that can describe the spatial pattern of a given phenotype is

$$y_i = f_{\mathrm{S}}(r(i), c(i)) + \varepsilon_i,\ \varepsilon_i \sim N(0, \sigma^2),\ i = 1, \ldots, M \tag{3.8}$$

where $f_S(r, c)$ is a smooth and unknown function that can be approximated by the tensor product of two one-dimensional cubic B-spline basis (one for each direction: row and column)

$$f_S(r,c) = \sum_{k_2=1}^{b_2}\sum_{k_3=1}^{b_3} B_{k_2}(r)B_{k_3}(c)\theta_{k_2k_3} = \sum_{k_2=1}^{b_2}\sum_{k_3=1}^{b_3} B_{k_2k_3}(r,c)\theta_{k_2k_3},$$

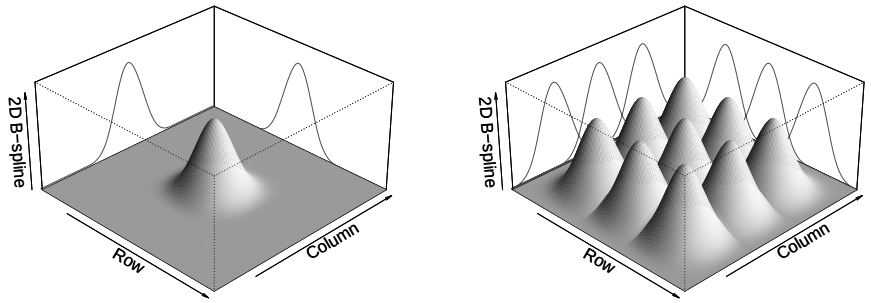
where $\boldsymbol{\theta}_S = (\theta_{11}, \ldots, \theta_{b_2 1}, \ldots, \theta_{b_2 b_3})^T$ is a vector of unknown regression coefficient. In matrix form, we have

$$\boldsymbol{f}_S = \boldsymbol{B}_S\boldsymbol{\theta}_S,$$

where $\boldsymbol{f}_S = (f_S(r(1), c(1)), \ldots, f_S(r(M), c(M)))^T$, and $\boldsymbol{B}_S$ is a (spatial) cubic B-spline design matrix such that

$$\boldsymbol{B}_S = \boldsymbol{B}_2 \square \boldsymbol{B}_3 = (\boldsymbol{B}_2 \otimes \mathbf{1}_{b_3}^T) \odot (\mathbf{1}_{b_2}^T \otimes \boldsymbol{B}_3), \tag{3.9}$$

where $\boldsymbol{B}_S$, of dimension $M \times b_2 b_3$, is calculated as the "box" product (the face-splitting product or row-wise Kronecker product, denoted as $\square$; Eilers et al., 2006; Slyusar, 1999) of two one-dimensional cubic B-spline design matrices in the row and column directions, $\boldsymbol{B}_2$ and $\boldsymbol{B}_3$, $\otimes$ denotes the Kronecker product, and $\odot$ the element-wise (Hadamard) product. We note that $(\boldsymbol{B}_2)_{ik_2}^{M\times b_2} = B_{k_2}(r(i))$, and $(\boldsymbol{B}_3)_{ik_3}^{M\times b_3} = B_{k_3}(c(i))$. For instance, Figure 3.7(a) is an illustrative representation of one tensor product of two one-dimensional cubic B-spline basis functions, $B_{k_2}(\boldsymbol{r})$ and $B_{k_3}(\boldsymbol{c})$. Figure 3.7(b) shows a portion (nine of them) of tensor products of two one-dimensional cubic B-spline basis functions.



(a) One tensor product (b) Nine tensor products

Figure 3.7: Illustrative visualisation of the tensor product of two one-dimensional cubic B-splines basis (in the row and column directions): (**a**) one tensor product of $B_{k_2}(\boldsymbol{r})$ and $B_{k_3}(\boldsymbol{c})$, (**b**) nine tensor products.

As for the one-dimensional case, in the two-dimensional setting smoothness is achieved by penalising coefficient differences, but now it has to be done along both the rows and columns. In particular, the penalised residual sum of squares is

$$\underbrace{\|\boldsymbol{y} - \boldsymbol{B}_S\boldsymbol{\theta}_S\|^2}_{\text{RSS}_S} + \underbrace{\boldsymbol{\theta}_S^T(\lambda_2\boldsymbol{I}_{b_3} \otimes \boldsymbol{P}_2 + \lambda_3\boldsymbol{P}_3 \otimes \boldsymbol{I}_{b_2})\boldsymbol{\theta}_S}_{\text{Spatial anisotropic penalty, } \boldsymbol{P}_S}, \tag{3.10}$$

where $\boldsymbol{P}_\nu = \boldsymbol{D}_\nu^T \boldsymbol{D}_\nu$, $\nu \in \{2, 3\}$, and $\boldsymbol{D}_\nu$ are second-order difference matrices. These difference penalties allow small discrepancies for adjacent coefficients within the same row or column (Eilers & Marx, 2003). We note that we are dealing with an anisotropic penalty in the sense that a different amount of smoothness on each dimension (λ_2 and λ_3) is allowed. The solution to the penalised residual sum of squares (3.10), given λ_2 and λ_3 is

$$\hat{\boldsymbol{\theta}}_{\mathrm{S}} = (\boldsymbol{B}_{\mathrm{S}}^T \boldsymbol{B}_{\mathrm{S}} + \boldsymbol{P}_{\mathrm{S}})^{-1} \boldsymbol{B}_{\mathrm{S}}^T \boldsymbol{y}, \tag{3.11}$$

where the size of the system of equations is $b_2 b_3$, the product of the number of columns of $\boldsymbol{B}_2$ and $\boldsymbol{B}_3$. As for the one-dimensional case, setting correct values for the smoothing parameters will be discussed in Section 3.2.

3.1.2.2 Three-dimensional P-splines: Spatio-temporal effects

Our motivation for the three-dimensional case is to model spatio-temporal effects simultaneously. The space is in two dimensions, rows and columns, and we need a third dimension for time. Then, $y_i(t)$ is the phenotype of interest for the ith plant ($i = 1, \ldots, M$) measured at time $t \in \{t_1, \ldots, t_n\}$. Plants are located in a rectangular grid (two-dimensional space) with R rows and C columns. As for the two-dimensional case, we use r and c for indices of row and column position, respectively ($r = 1, \ldots, R$; $c = 1, \ldots, C$), and $r(i)$ and $c(i)$ to denote the row and column position of the ith plant, respectively. Note that we assume that all plants in the experiment are measured at the same times, but missing data is allowed. For instance, we are interested in studying the spatio-temporal effects of the leaf area for the PhenoArch platform, as depicted in Figure 2.7. As a first approach, the three-dimensional interaction model is

$$y_i(t) = f_{\mathrm{ST}}(r(i), c(i), t) + \varepsilon_i(t),\ \varepsilon_i(t) \sim N(0, \sigma^2),\ i = 1, \ldots, M,\ t \in \{t_1, \ldots, t_n\}, \tag{3.12}$$

where $f_{\mathrm{ST}}(r, c, t)$ is a smooth and unknown three-dimensional function that considers the spatio-temporal interaction. Following the ideas in Lee and Durban (2011), we approximate this function by using the tensor-product of three one-dimensional cubic B-splines basis, i.e.,

$$f_{\mathrm{ST}}(r, c, t) = \sum_{k_2=1}^{b_2} \sum_{k_3=1}^{b_3} \sum_{k_1=1}^{b_1} B_{k_2}(r) B_{k_3}(c) B_{k_1}(t) \theta_{k_2 k_3 k_1} = \sum_{k_2=1}^{b_2} \sum_{k_3=1}^{b_3} \sum_{k_1=1}^{b_1} B_{k_2 k_3 k_1}(r, c, t) \theta_{k_2 k_3 k_1}, \tag{3.13}$$

where $\boldsymbol{f}_{\mathrm{ST}} = (f_{\mathrm{ST}}(r(1), c(1), t_1), \ldots, f_{\mathrm{ST}}(r(1), c(1), t_n), \ldots, f_{\mathrm{ST}}(r(M), c(M), t_n))^T = \boldsymbol{B}_{\mathrm{ST}} \boldsymbol{\theta}_{\mathrm{ST}}$, with $\boldsymbol{\theta}_{\mathrm{ST}} = (\theta_{111}, \ldots, \theta_{11b_1}, \theta_{121}, \ldots, \theta_{12b_1}, \ldots, \theta_{1b_31}, \ldots, \theta_{1b_3b_1}, \theta_{211}, \ldots, \theta_{21b_1}, \ldots, \theta_{b_2b_31}, \ldots, \theta_{b_2b_3b_1})^T$ the vector of coefficients, and the compound spatio-temporal B-spline design matrix given by

$$\boldsymbol{B}_{\mathrm{ST}} = \boldsymbol{B}_{\mathrm{S}} \otimes \boldsymbol{B}_{\mathrm{T}}, \tag{3.14}$$

where $\boldsymbol{B}_{\mathrm{ST}}$ of dimension $Mn \times b_2 b_3 b_1$ is the Kronecker product of the bi-dimensional spatial cubic B-spline design matrix defined in (3.9), i.e., $\boldsymbol{B}_{\mathrm{S}}^{M \times b_2 b_3}$, and the unidimensional temporal cubic B-spline design matrix

defined in (3.3), i.e., $\boldsymbol{B}_{\mathrm{T}}^{n\times b_1}$. As for the bi-dimensional case, smoothness along rows, columns and time is imposed by an anisotropic penalty such that the penalised residual sum of squares is

$$\underbrace{\|\boldsymbol{y}-\boldsymbol{B}_{\mathrm{ST}}\boldsymbol{\theta}_{\mathrm{ST}}\|^2}_{\mathrm{RSS}_{\mathrm{ST}}}+\boldsymbol{\theta}_{\mathrm{ST}}^T\underbrace{(\lambda_2\boldsymbol{P}_2\otimes\boldsymbol{I}_{b_3}\otimes\boldsymbol{I}_{b_1}+\lambda_3\boldsymbol{I}_{b_2}\otimes\boldsymbol{P}_3\otimes\boldsymbol{I}_{b_1}+\lambda_1\boldsymbol{I}_{b_2}\otimes\boldsymbol{I}_{b_3}\otimes\boldsymbol{P}_{\mathrm{T}}^T)}_{\text{Spatio-temporal anisotropic penalty, }\boldsymbol{P}_{\mathrm{ST}}}\boldsymbol{\theta}_{\mathrm{ST}}, \tag{3.15}$$

where $\boldsymbol{P}_\nu=\boldsymbol{D}_\nu^T\boldsymbol{D}_\nu$, $\nu\in\{1,2,3\}$) are based on second-order difference matrices, and λ_1, λ_2 and λ_3 are the smoothing parameters. The explicit solution of the penalised objective function (3.15), given λ_1, λ_2 and λ_3 is

$$\hat{\boldsymbol{\theta}}_{ST}=(\boldsymbol{B}_{\mathrm{ST}}^T\boldsymbol{B}_{\mathrm{ST}}+\boldsymbol{P}_{\mathrm{ST}})^{-1}\boldsymbol{B}_{\mathrm{ST}}^T\boldsymbol{y}, \tag{3.16}$$

where the size of the system of equations is $b_2b_3b_1$. All in all, we have seen how the complexity of the solution increases with the dimensionality of the problem. In one dimension, it is easy to be generous with the number of cubic B-spline basis used. For instance, if we use cubic B-spline basis of size $b_1=27$, we will obtain 27 coefficients. For two dimensions, if we use a tensor product of two cubic B-spline basis with, e.g., $b_2=b_3=27$, we will obtain $b_2\times b_3=729$ coefficients. If we additionally have a third dimension, this yields $b_2\times b_3\times b_1=19683$ coefficients. Alternatively, to obtain 729 coefficients (as in the bi-dimensional problem), we would need cubic B-spline basis of size 9 in each direction. Thus, the size of the cubic B-spline basis together with the data size (a large number of plants and time points), could make the estimation problem computationally challenging. We discuss this point in detail in Section 3.3.

3.2 P-splines and mixed model formulation

In the previous section, we obtained closed-form expressions for the estimates of the vector of regression coefficients (see expressions (3.6), (3.11) and (3.16)) for given smoothing parameters. However, when working with penalised regression, we need to find appropriate values for these smoothing parameters (λ_ν, $\nu\in\{1,2,3\}$), since the results will heavily depend on this choice. To tackle this problem, one option is to use the connection between P-splines and mixed models through the parameterisation proposed by Currie and Durban (2002) and Wand (2003). In this parameterisation, the smooth functions are sums of fixed (unpenalised) and random (penalised) components, and the smoothing parameters are replaced by ratios of variances, which are estimated by restricted maximum likelihood (REML; Patterson & Thompson, 1971). As before, we introduce mixed model formulation in one-dimension and we later extend the ideas for the multidimensional case (two- and three-dimensions).

3.2.1 Mixed model formulation of P-splines in one dimension

For the one-dimensional model (3.2), we note that, for a given smoothing parameter, λ_1, the solution to the penalised residual sum of squares in (3.5) corresponds to the empirical best linear unbiased predictors

(BLUP) for $\boldsymbol{\theta}_{\mathrm{T}}$ under the assumption that $\boldsymbol{\theta}_{\mathrm{T}} \sim N(\mathbf{0}, \sigma^2/\lambda_1 \boldsymbol{P}_{\mathrm{T}}^{-})$, where $\boldsymbol{P}_{\mathrm{T}}^{-}$ denotes the Moore-Penrose inverse of the penalty matrix $\boldsymbol{P}_{\mathrm{T}} = \boldsymbol{D}_1^T \boldsymbol{D}_1$ (we need to work with the generalised inverse since $\boldsymbol{P}_{\mathrm{T}}$ is not of full rank). To obtain a full rank precision matrix (i.e., the inverse of the variance-covariance matrix), we write the P-spline model as the following standard linear mixed model

$$\boldsymbol{y} = \boldsymbol{B}_{\mathrm{T}}\boldsymbol{\theta}_{\mathrm{T}} + \boldsymbol{\varepsilon} = \boldsymbol{X}_{\mathrm{T}}\boldsymbol{\beta}_{\mathrm{T}} + \boldsymbol{Z}_{\mathrm{T}}\boldsymbol{u}_{\mathrm{T}} + \boldsymbol{\varepsilon}; \quad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{R}_{\mathrm{T}}), \quad \boldsymbol{u}_{\mathrm{T}} \sim N(\mathbf{0}, \boldsymbol{G}_{\mathrm{T}}), \tag{3.17}$$

where $\boldsymbol{X}_{\mathrm{T}}$ (with $d = 2$ columns) and $\boldsymbol{Z}_{\mathrm{T}}$ (with $b_1 - 2$ columns) are the mixed model design matrices, $\boldsymbol{\beta}_{\mathrm{T}}$ and $\boldsymbol{u}_{\mathrm{T}}$ are the vectors of fixed and random coefficients, respectively, $\boldsymbol{R}_{\mathrm{T}} = \sigma^2 \boldsymbol{I}_n$, and $\boldsymbol{G}_{\mathrm{T}}$ is the diagonal variance-covariance matrix for the random effects. To decompose $\boldsymbol{B}_{\mathrm{T}}$ in the proposed way, we follow Currie et al. (2006) and Lee and Durban (2011), and use the singular value decomposition (SVD) of $\boldsymbol{P}_{\mathrm{T}} = \boldsymbol{D}_1^T \boldsymbol{D}_1 = \boldsymbol{U}_{\mathrm{T}}\boldsymbol{\Lambda}_{\mathrm{T}}\boldsymbol{U}_{\mathrm{T}}^T$. Here $\boldsymbol{U}_{\mathrm{T}}$ is the matrix of eigenvectors and $\boldsymbol{\Lambda}_{\mathrm{T}}$ is the diagonal matrix of eigenvalues. Let us also denote $\boldsymbol{U}_{\mathrm{T}+}$ ($\boldsymbol{\Lambda}_{\mathrm{T}+}$) and $\boldsymbol{U}_{\mathrm{T}0}$ ($\boldsymbol{\Lambda}_{\mathrm{T}0}$) the sub-matrices corresponding to the non-zero and zero eigenvalues, respectively (we note that for second-order difference penalties, there are two zero eigenvalues, and thus $\boldsymbol{\Lambda}_{\mathrm{T}0} = \mathbf{0}_2$ is a 2-by-2 matrix of zeroes). Given that $\boldsymbol{U}_{\mathrm{T}}\boldsymbol{U}_{\mathrm{T}}^T = \boldsymbol{I}_{b_1}$, it is easy to show that the P-spline model can be parameterised as

$$\begin{aligned}
\boldsymbol{y} &= \boldsymbol{B}_{\mathrm{T}}\boldsymbol{U}_{\mathrm{T}}\boldsymbol{U}_{\mathrm{T}}^T\boldsymbol{\theta}_{\mathrm{T}} + \boldsymbol{\varepsilon} \\
&= \boldsymbol{B}_{\mathrm{T}}[\boldsymbol{U}_{\mathrm{T}0}|\boldsymbol{U}_{\mathrm{T}+}][\boldsymbol{U}_{\mathrm{T}0}^T|\boldsymbol{U}_{\mathrm{T}+}^T]\boldsymbol{\theta}_{\mathrm{T}} + \boldsymbol{\varepsilon} \\
&= \boldsymbol{B}_{\mathrm{T}}\boldsymbol{U}_{\mathrm{T}0}\boldsymbol{U}_{\mathrm{T}0}^T\boldsymbol{\theta}_{\mathrm{T}} + \boldsymbol{B}_{\mathrm{T}}\boldsymbol{U}_{\mathrm{T}+}\boldsymbol{U}_{\mathrm{T}+}^T\boldsymbol{\theta}_{\mathrm{T}} + \boldsymbol{\varepsilon}.
\end{aligned} \tag{3.18}$$

As such, $\boldsymbol{X}_{\mathrm{T}} = \boldsymbol{B}_{\mathrm{T}}\boldsymbol{U}_{\mathrm{T}0}$ and $\boldsymbol{Z}_{\mathrm{T}} = \boldsymbol{B}_{\mathrm{T}}\boldsymbol{U}_{\mathrm{T}+}$, and $\boldsymbol{\beta}_{\mathrm{T}} = \boldsymbol{U}_{\mathrm{T}0}^T\boldsymbol{\theta}_{\mathrm{T}}$ and $\boldsymbol{u}_{\mathrm{T}} = \boldsymbol{U}_{\mathrm{T}+}^T\boldsymbol{\theta}_{\mathrm{T}}$ (i.e., $\boldsymbol{\theta}_{\mathrm{T}} = \boldsymbol{U}_{\mathrm{T}0}\boldsymbol{\beta}_{\mathrm{T}} + \boldsymbol{U}_{\mathrm{T}+}\boldsymbol{u}_{\mathrm{T}}$). It follows that the penalty term in (3.5) can also be rewritten as

$$\begin{aligned}
\lambda_1\boldsymbol{\theta}_{\mathrm{T}}^T\boldsymbol{P}_{\mathrm{T}}\boldsymbol{\theta}_{\mathrm{T}} &= \lambda_1\boldsymbol{\theta}_{\mathrm{T}}^T[\boldsymbol{U}_{\mathrm{T}0}|\boldsymbol{U}_{\mathrm{T}+}]\begin{bmatrix}\mathbf{0}_2 & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Lambda}_{\mathrm{T}+}\end{bmatrix}[\boldsymbol{U}_{\mathrm{T}0}^T|\boldsymbol{U}_{\mathrm{T}+}^T]\boldsymbol{\theta}_{\mathrm{T}} \\
&= \lambda_1[\boldsymbol{\theta}_{\mathrm{T}}^T\boldsymbol{U}_{\mathrm{T}0}|\boldsymbol{\theta}_{\mathrm{T}}^T\boldsymbol{U}_{\mathrm{T}+}]\begin{bmatrix}\mathbf{0}_2 & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Lambda}_{\mathrm{T}+}\end{bmatrix}[\boldsymbol{U}_{\mathrm{T}}^T\boldsymbol{\theta}_{\mathrm{T}}|\boldsymbol{U}_{\mathrm{T}+}^T\boldsymbol{\theta}_{\mathrm{T}}] \\
&= \lambda_1[\boldsymbol{\beta}_{\mathrm{T}}^T|\boldsymbol{u}_{\mathrm{T}}^T]\begin{bmatrix}\mathbf{0}_2 & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Lambda}_{\mathrm{T}+}\end{bmatrix}[\boldsymbol{\beta}_{\mathrm{T}}|\boldsymbol{u}_{\mathrm{T}}] \\
&= \lambda_1\boldsymbol{u}_{\mathrm{T}}^T\boldsymbol{\Lambda}_{\mathrm{T}+}\boldsymbol{u}_{\mathrm{T}}.
\end{aligned}$$

That is to say, only $\boldsymbol{u}_{\mathrm{T}}$ is penalised, while $\boldsymbol{\beta}_{\mathrm{T}}$ is not. In the equivalent linear mixed model, it implies that $\boldsymbol{\beta}_{\mathrm{T}} = (\beta_0, \beta_1)^T$ is a vector of unpenalised/fixed effects, and $\boldsymbol{u}_{\mathrm{T}} = (u_1, \ldots, u_{b_1-2})^T$ is a vector of penalised/random effects, assumed to be distributed according to $\sim N(\mathbf{0}, \boldsymbol{G}_{\mathrm{T}})$, where $\boldsymbol{G}_{\mathrm{T}} = \sigma_1^2\boldsymbol{\Sigma}_{\mathrm{T}+}$, with $\sigma_1^2 = \sigma^2/\lambda_1$ and $\boldsymbol{\Sigma}_{\mathrm{T}+} = \boldsymbol{\Lambda}_{\mathrm{T}+}^{-1}$. That is, the smoothing parameter, λ_1, is now the ratio between two variance parameters: the residual variance, σ^2, and the variance of the random effects, σ_1^2. Then, the penalised residual sum of squares (3.5) can be reformulated as

$$\begin{aligned}
&\|\boldsymbol{y} - \boldsymbol{X}_{\mathrm{T}}\boldsymbol{\beta}_{\mathrm{T}} - \boldsymbol{Z}_{\mathrm{T}}\boldsymbol{u}_{\mathrm{T}}\|^2 + \lambda_1\boldsymbol{u}_{\mathrm{T}}^T\boldsymbol{\Lambda}_{\mathrm{T}+}\boldsymbol{u}_{\mathrm{T}} \\
&\|\boldsymbol{y} - \boldsymbol{X}_{\mathrm{T}}\boldsymbol{\beta}_{\mathrm{T}} - \boldsymbol{Z}_{\mathrm{T}}\boldsymbol{u}_{\mathrm{T}}\|^2 + \sigma^2\boldsymbol{u}_{\mathrm{T}}^T\boldsymbol{G}_{\mathrm{T}}^{-1}\boldsymbol{u}_{\mathrm{T}}.
\end{aligned}$$

We note that based on the SVD, $\boldsymbol{X}_{\mathrm{T}} = \boldsymbol{B}_{\mathrm{T}}\boldsymbol{U}_{\mathrm{T0}}$. However, it is sometimes more convenient (or, at least, it helps to understand the unpenalised/fixed part of a P-spline) to take $\boldsymbol{X}_{\mathrm{T}} = [\mathbf{1}_n \mid \boldsymbol{t}]$. In other words, when using P-splines in combination with a second-order penalty, the space of functions that are not penalised corresponds to the polynomials of degree 1. Another possible parameterisation for $\boldsymbol{X}_{\mathrm{T}}$ is the one proposed by Wood et al. (2013), in which we also obtain a design matrix with a constant column (no necessarily of ones). For this purpose, we first obtain $\tilde{\boldsymbol{X}}_{\mathrm{T}} = \boldsymbol{B}_{\mathrm{T}}\boldsymbol{U}_{\mathrm{T0}}$ as before, and then compute the singular value decomposition of $\boldsymbol{F}_{\mathrm{T}}^T\boldsymbol{F}_{\mathrm{T}} = \boldsymbol{V}_{\mathrm{T}}\boldsymbol{\Omega}_{\mathrm{T}}\boldsymbol{V}_{\mathrm{T}}^T$, with $\boldsymbol{F}_{\mathrm{T}} = \tilde{\boldsymbol{X}}_{\mathrm{T}} - \mathbf{1}\mathbf{1}^T\tilde{\boldsymbol{X}}_{\mathrm{T}}/n$ (i.e., based on centered values for $\tilde{\boldsymbol{X}}_{\mathrm{T}}$), and then define $\boldsymbol{X}_{\mathrm{T}} = \boldsymbol{B}_{\mathrm{T}}\boldsymbol{U}_{\mathrm{T0}}\boldsymbol{V}_{\mathrm{T}}$. For simplicity, we use the parameterisation $\boldsymbol{X}_{\mathrm{T}} = [\mathbf{1}_n \mid \boldsymbol{t}]$ to explicitly understand the expression of the P-spline (3.2) into the mixed model formulation

$$f_{\mathrm{T}}(t) = \beta_0 + \beta_1 t + \sum_{k_1=1}^{b_t-2} z_{k_1}(t)u_{k_1} \tag{3.19}$$

where $\{z_{k_1}(\cdot) : 1 \leq k_1 \leq b_1 - 2\}$ is the set of basis functions obtained from the connection between P-splines and linear mixed models.

3.2.2 Mixed model formulation of P-splines in two dimensions

We now turn in the mixed model formulation of the bi-dimensional P-spline model. We present here the main ideas, but more details can be found in Eilers et al. (2006), Lee (2010), Lee and Durban (2011), and Lee et al. (2013). As for the one-dimensional case, we need to obtain the mixed model matrices $\boldsymbol{X}_{\mathrm{S}}$ and $\boldsymbol{Z}_{\mathrm{S}}$, the vectors of regression coefficients $\boldsymbol{\beta}_{\mathrm{S}}$ and $\boldsymbol{u}_{\mathrm{S}}$, and the variance-covariance matrix $\boldsymbol{G}_{\mathrm{S}}$, such that

$$\boldsymbol{f}_{\mathrm{S}} = \boldsymbol{B}_{\mathrm{S}}\boldsymbol{\theta}_{\mathrm{S}} = \boldsymbol{X}_{\mathrm{S}}\boldsymbol{\beta}_{\mathrm{S}} + \boldsymbol{Z}_{\mathrm{S}}\boldsymbol{u}_{\mathrm{S}}, \quad \boldsymbol{u}_{\mathrm{S}} \sim N(\mathbf{0}, \boldsymbol{G}_{\mathrm{S}}), \tag{3.20}$$

where a diagonal variance-covariance matrix $\boldsymbol{G}_{\mathrm{S}}$ can be obtained by simultaneously diagonalising the marginal penalties $\boldsymbol{P}_2$ and $\boldsymbol{P}_3$ in (3.10). To that aim, we construct a compound (spatial) transformation matrix, $\boldsymbol{U}_{\mathrm{S}}$, on the basis of the transformation matrix for each marginal penalties. In particular, let $\boldsymbol{P}_\nu = \boldsymbol{D}_\nu^T\boldsymbol{D}_\nu = \boldsymbol{U}_\nu\boldsymbol{\Lambda}_\nu\boldsymbol{U}_\nu^T$ be the SVD of $\boldsymbol{D}_\nu^T\boldsymbol{D}_\nu$, $\boldsymbol{U}_\nu = [\boldsymbol{U}_{\nu 0}|\boldsymbol{U}_{\nu +}]$ ($\nu \in \{2, 3\}$), and define $\boldsymbol{U}_{\mathrm{S}}$ as

$$\begin{aligned}
\boldsymbol{U}_{\mathrm{S}} &= \boldsymbol{U}_2 \otimes \boldsymbol{U}_3 \\
&= [\boldsymbol{U}_{20}|\boldsymbol{U}_{2+}] \otimes [\boldsymbol{U}_{30}|\boldsymbol{U}_{3+}] \\
&= [\underbrace{\boldsymbol{U}_{20} \otimes \boldsymbol{U}_{30}}_{\boldsymbol{U}_{\mathrm{S0}}} \mid \underbrace{\boldsymbol{U}_{20} \otimes \boldsymbol{U}_{3+}|\boldsymbol{U}_{2+} \otimes \boldsymbol{U}_{30}|\boldsymbol{U}_{2+} \otimes \boldsymbol{U}_{3+}}_{\boldsymbol{U}_{\mathrm{S}+}}].
\end{aligned} \tag{3.21}$$

Accordingly, the mixed model design matrices $\boldsymbol{X}_{\mathrm{S}}$ and $\boldsymbol{Z}_{\mathrm{S}}$ are obtained by multiplying the spatial cubic B-spline design matrix $\boldsymbol{B}_{\mathrm{S}}$ in (3.9) and the appropriate component of the transformation matrix $\boldsymbol{U}_{\mathrm{S}}$ in (3.21)

$$
\begin{aligned}
\boldsymbol{X}_{\mathrm{S}} &= (\boldsymbol{B}_2 \square \boldsymbol{B}_3)\boldsymbol{U}_{\mathrm{S}0} \\
&= (\boldsymbol{B}_2 \square \boldsymbol{B}_3)[\boldsymbol{U}_{20} \otimes \boldsymbol{U}_{30}] \\
&= [\boldsymbol{B}_2\boldsymbol{U}_{20} \square \boldsymbol{B}_3\boldsymbol{U}_{30}] \\
&= [\boldsymbol{X}_2 \square \boldsymbol{X}_3] \\
\boldsymbol{Z}_{\mathrm{S}} &= (\boldsymbol{B}_2 \square \boldsymbol{B}_3)\boldsymbol{U}_{\mathrm{S}+} \\
&= (\boldsymbol{B}_2 \square \boldsymbol{B}_3)[\boldsymbol{U}_{20} \otimes \boldsymbol{U}_{3+} | \boldsymbol{U}_{2+} \otimes \boldsymbol{U}_{30} | \boldsymbol{U}_{2+} \otimes \boldsymbol{U}_{3+}] \\
&= [\boldsymbol{B}_2\boldsymbol{U}_{20} \square \boldsymbol{B}_3\boldsymbol{U}_{3+} | \boldsymbol{B}_2\boldsymbol{U}_{2+} \square \boldsymbol{B}_3\boldsymbol{U}_{30} | \boldsymbol{B}_2\boldsymbol{U}_{2+} \square \boldsymbol{B}_3\boldsymbol{U}_{3+}] \\
&= [\boldsymbol{X}_2 \square \boldsymbol{Z}_3 | \boldsymbol{Z}_2 \square \boldsymbol{X}_3 | \boldsymbol{Z}_2 \square \boldsymbol{Z}_3],
\end{aligned}
\tag{3.22}
$$

where $\boldsymbol{X}_2 = \boldsymbol{B}_2\boldsymbol{U}_{20}$, $\boldsymbol{X}_3 = \boldsymbol{B}_3\boldsymbol{U}_{30}$, $\boldsymbol{Z}_2 = \boldsymbol{B}_2\boldsymbol{U}_{2+}$, $\boldsymbol{Z}_3 = \boldsymbol{B}_3\boldsymbol{U}_{3+}$ and $\boldsymbol{\beta}_{\mathrm{S}} = \boldsymbol{\theta}_{\mathrm{S}}\boldsymbol{U}_{\mathrm{S}0}^T$ and $\boldsymbol{u}_{\mathrm{S}} = \boldsymbol{\theta}_{\mathrm{S}}\boldsymbol{U}_{\mathrm{S}+}^T$. Likewise, the spatial anisotropic penalty $\boldsymbol{P}_{\mathrm{S}}$ in (3.10) can be specified as

$$
\begin{aligned}
\boldsymbol{U}_{\mathrm{S}}^T\boldsymbol{P}_{\mathrm{S}}\boldsymbol{U}_{\mathrm{S}} &= \boldsymbol{U}_{\mathrm{S}}^T(\lambda_2\boldsymbol{I}_{b_3} \otimes \boldsymbol{P}_2 + \lambda_3\boldsymbol{P}_3 \otimes \boldsymbol{I}_{b_2})\boldsymbol{U}_{\mathrm{S}} \\
&= \boldsymbol{U}_{\mathrm{S}}^T(\lambda_2\boldsymbol{I}_{b_3} \otimes \boldsymbol{P}_2)\boldsymbol{U}_{\mathrm{S}} + \boldsymbol{U}_{\mathrm{S}}^T(\lambda_3\boldsymbol{P}_3 \otimes \boldsymbol{I}_{b_2})\boldsymbol{U}_{\mathrm{S}},
\end{aligned}
$$

where

$$
\boldsymbol{U}_{\mathrm{S}}^T(\lambda_2\boldsymbol{I}_{b_3} \otimes \boldsymbol{P}_2)\boldsymbol{U}_{\mathrm{S}} = \lambda_2 \begin{pmatrix} \boldsymbol{0}_{2\times 2} & & \\ & \boldsymbol{I}_2 \otimes \boldsymbol{\Lambda}_{2+} & \\ & & \boldsymbol{0}_{2(b_3-2)} \\ & & \boldsymbol{I}_{b_3-2} \otimes \boldsymbol{\Lambda}_{2+} \end{pmatrix},
$$

and

$$
\boldsymbol{U}_{\mathrm{S}}^T(\lambda_3\boldsymbol{P}_3 \otimes \boldsymbol{I}_{b_2})\boldsymbol{U}_{\mathrm{S}} = \lambda_3 \begin{pmatrix} \boldsymbol{0}_{2\times 2} & & \\ & \boldsymbol{0}_{2(b_2-2)} & \\ & & \boldsymbol{\Lambda}_{3+} \otimes \boldsymbol{I}_2 \\ & & \boldsymbol{\Lambda}_{3+} \otimes \boldsymbol{I}_{b_2-2} \end{pmatrix}.
$$

We note that the first block in both matrices corresponds to the unpenalised (fixed) coefficients of the model. The other blocks correspond to the precision matrix $\boldsymbol{G}_{\mathrm{S}}^{-1}$ of the penalised (random) coefficients, which in this case is

$$
\begin{aligned}
\boldsymbol{G}_{\mathrm{S}}^{-1} &= \begin{pmatrix} \frac{1}{\sigma_2^2}\boldsymbol{I}_2 \otimes \boldsymbol{\Lambda}_{2+} & & \\ & \frac{1}{\sigma_3^2}\boldsymbol{\Lambda}_{3+} \otimes \boldsymbol{I}_2 & \\ & & \frac{1}{\sigma_2^2}\boldsymbol{I}_{b_3-2} \otimes \boldsymbol{\Lambda}_{2+} + \frac{1}{\sigma_3^2}\boldsymbol{\Lambda}_{3+} \otimes \boldsymbol{I}_{b_2-2} \end{pmatrix} \\
&= \frac{1}{\sigma_2^2}\underbrace{\begin{pmatrix} \boldsymbol{I}_2 \otimes \boldsymbol{\Lambda}_{2+} & & \\ & \boldsymbol{0}_{2(b_3-2)} & \\ & & \boldsymbol{I}_{b_3-2} \otimes \boldsymbol{\Lambda}_{2+} \end{pmatrix}}_{\tilde{\boldsymbol{\Lambda}}_{2+}} + \frac{1}{\sigma_3^2}\underbrace{\begin{pmatrix} \boldsymbol{0}_{2(b_2-2)} & & \\ & \boldsymbol{\Lambda}_{3+} \otimes \boldsymbol{I}_2 & \\ & & \boldsymbol{\Lambda}_{3+} \otimes \boldsymbol{I}_{b_2-2} \end{pmatrix}}_{\tilde{\boldsymbol{\Lambda}}_{3+}} \\
&= \sum_{j\in\{2,3\}} \sigma_j^{-2}\tilde{\boldsymbol{\Lambda}}_{j+},
\end{aligned}
\tag{3.23}
$$

where $\sigma_1^2 = \sigma^2/\lambda_1$ and $\sigma_2^2 = \sigma^2/\lambda_2$. Note that the last block in $\boldsymbol{G}_{\mathrm{S}}^{-1}$ involves both variance parameters σ_2^2 and σ_3^2. If we compute the variance-covariance matrix, $\boldsymbol{G}_{\mathrm{S}}$, we will have a block involving both variance parameters but in a non-linear way. Consequently, standard mixed model estimation techniques can not be used. As we will discuss in Section 3.3, to estimate the variance parameters, we will take advantage of the fact that the precision matrix $\boldsymbol{G}_{\mathrm{S}}^{-1}$ in (3.23) is linear on the inverse of the variance parameters, i.e. it is linear on the precision parameters σ_2^{-2} and σ_3^{-2}. Another option for estimating the variance parameters is to reorganize the bivariate P-spline model and use the PS-ANOVA model proposed by Lee et al. (2013). Here, we will assume that $\boldsymbol{X}_2 = [\mathbf{1}_M|\boldsymbol{r}]$ and $\boldsymbol{X}_3 = [\mathbf{1}_M|\boldsymbol{c}]$ (or more precisely the parameterisation proposed by Wood et al., 2013). We then can construct the compound mixed model matrices as follows

$$\begin{aligned}\boldsymbol{X}_{\mathrm{S}} &= [\boldsymbol{X}_3 \square \boldsymbol{X}_2] \equiv [\mathbf{1}_M \square \mathbf{1}_M | \boldsymbol{c} \square \mathbf{1}_M | \mathbf{1}_M \square \boldsymbol{r} | \boldsymbol{r} \square \boldsymbol{c}] = [\mathbf{1}_M | \boldsymbol{c} | \boldsymbol{r} | \boldsymbol{r} \odot \boldsymbol{c}] \\ \boldsymbol{Z}_{\mathrm{S}} &= [\boldsymbol{X}_3 \square \boldsymbol{Z}_2 | \boldsymbol{Z}_3 \square \boldsymbol{X}_2 | \boldsymbol{Z}_3 \square \boldsymbol{Z}_2] \equiv [\boldsymbol{Z}_3 \square \mathbf{1}_M | \mathbf{1}_M \square \boldsymbol{Z}_2 | \boldsymbol{Z}_3 \square \boldsymbol{r} | \boldsymbol{c} \square \boldsymbol{Z}_2 | \boldsymbol{Z}_3 \square \boldsymbol{Z}_2] = [\boldsymbol{Z}_3 | \boldsymbol{Z}_2 | \boldsymbol{Z}_3 \square \boldsymbol{r} | \boldsymbol{c} \square \boldsymbol{Z}_2 | \boldsymbol{Z}_3 \square \boldsymbol{Z}_2],\end{aligned} \tag{3.24}$$

where $\equiv$ denotes that previous and actual matrices have the same elements, but in different order, and $\odot$ denotes the element-wise vector (matrix) product. Note that the matrix of random effects, $\boldsymbol{Z}_S$ has five blocks. It then follows that the precision matrix has five blocks, one associated with each term in $\boldsymbol{Z}_S$, as follows

$$\boldsymbol{G}_{\mathrm{S}}^{-1} = \begin{pmatrix} \frac{1}{\sigma_3^2}\boldsymbol{\Lambda}_{3+} & & & & & \\ & \frac{1}{\sigma_2^2}\boldsymbol{\Lambda}_{2+} & & & & \\ & & \frac{1}{\sigma_3^2}\boldsymbol{\Lambda}_{3+} & & & \\ & & & \frac{1}{\sigma_2^2}\boldsymbol{\Lambda}_{2+} & & \\ & & & & \frac{1}{\sigma_2^2}\boldsymbol{I}_{b_2-2} \otimes \boldsymbol{\Lambda}_{2+} + \frac{1}{\sigma_3^2}\boldsymbol{\Lambda}_{3+} \otimes \boldsymbol{I}_{b_3-2} \end{pmatrix}. \tag{3.25}$$

As discussed in Lee et al. (2013), the block structure of both $\boldsymbol{X}_{\mathrm{S}}$ and $\boldsymbol{Z}_{\mathrm{S}}$ (see (3.24)) results in the following interesting ANOVA-type decomposition of the bivariate smooth surface

$$\boldsymbol{f}_{\mathrm{S}} = \underbrace{\mathbf{1}_M\beta_0 + \boldsymbol{c}\beta_1 + \boldsymbol{r}\beta_2 + (\boldsymbol{r} \odot \boldsymbol{c})\beta_3}_{\text{Bilinear polynomial } (\boldsymbol{X}_S\boldsymbol{\beta}_S)} + \underbrace{\underbrace{h_3(\boldsymbol{c})}_{\boldsymbol{Z}_3\boldsymbol{u}_1} + \underbrace{h_2(\boldsymbol{r})}_{\boldsymbol{Z}_2\boldsymbol{u}_2} + \underbrace{\boldsymbol{r} \odot f_{3;2}(\boldsymbol{c})}_{(\boldsymbol{Z}_3 \square \boldsymbol{r})\boldsymbol{u}_3} + \underbrace{\boldsymbol{c} \odot f_{2;3}(\boldsymbol{r})}_{(\boldsymbol{c} \square \boldsymbol{Z}_2)\boldsymbol{u}_4} + \underbrace{f_{2|3}(\boldsymbol{r}, \boldsymbol{c})}_{(\boldsymbol{Z}_3 \square \boldsymbol{Z}_2)\boldsymbol{u}_5}}_{\text{Smooth term } (\boldsymbol{Z}_S\boldsymbol{u}_S)}, \tag{3.26}$$

where $\boldsymbol{u}_k$ ($k = 1, \ldots, 5$) contains the elements of $\boldsymbol{u}_S$ that correspond to the kth block of $\boldsymbol{Z}_S$. The interpretation of (3.26) is as follows. The (unpenalised/fixed) bilinear term contains the intercept (β_0), the linear trends along the column (β_1) and row (β_2) directions, as well as the linear interaction trend (β_3). The (penalised/random) smooth term includes the smooth (non-linear) trends (main effects) along columns, $h_3(c)$, and rows, $h_2(r)$; $r \times f_{3;2}(c)$ and $c \times f_{2;3}(r)$ are linear-by-smooth interaction trends (varying coefficient surface terms; for instance $c \times f_{2;3}(r)$ are linear trends in the columns (c) but with slopes ($f_{2;3}(r)$) that change

smoothly along the rows); and the pure smooth-by-smooth interaction trend jointly defined over the row and column directions ($f_{2|3}(r, c)$). A decomposition of this type is depicted in Figure 3.8.

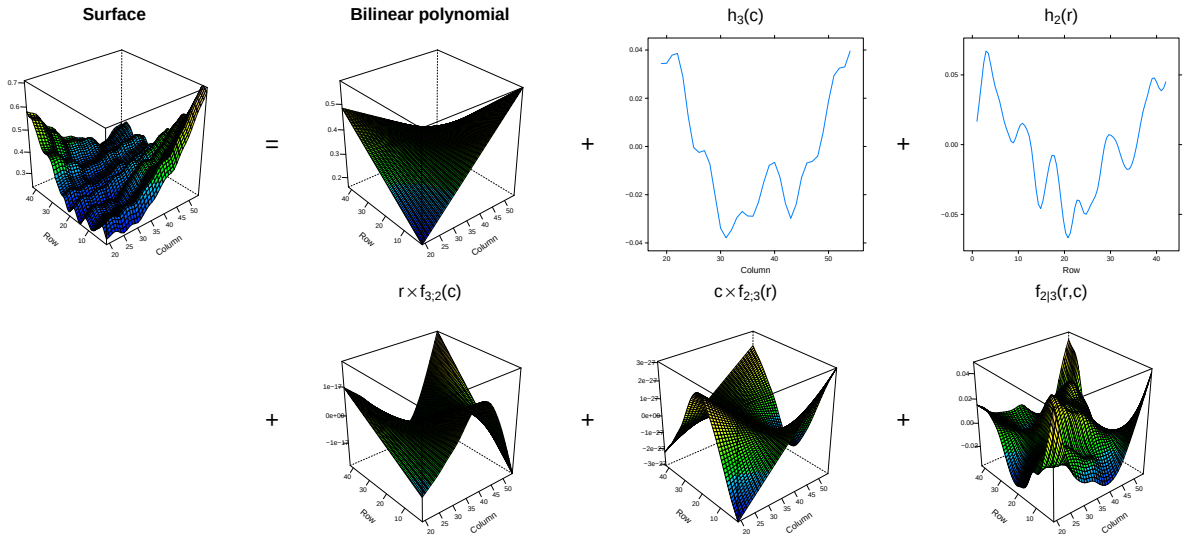


Figure 3.8: For the FIP platform data (trial 2017) described in Section 2.2, smooth components of the ANOVA-type decomposition (see equation (3.26)) of the estimated spatial trend for one time point.

A close look at (3.25) shows that, despite the five smooth components in (3.8), only two variance parameters (or smoothing parameters) control their smoothness (σ_2^2 and σ_3^2). In fact, the same variance parameters apply to both, main effects and interaction terms. In Lee et al. (2013) the ANOVA-type decomposition is further exploited, and a different variance parameter is considered for each smooth term, i.e., each block in (3.25) will have its own variance parameters. For ease of notation, let $\boldsymbol{\Lambda}_1 = \boldsymbol{\Lambda}_3 = \boldsymbol{\Lambda}_{3+}$, $\boldsymbol{\Lambda}_2 = \boldsymbol{\Lambda}_4 = \boldsymbol{\Lambda}_{2+}$, and $\boldsymbol{\Lambda}_5 = \boldsymbol{I}_{b_2-2} \otimes \boldsymbol{\Lambda}_{2+} + \boldsymbol{\Lambda}_{3+} \otimes \boldsymbol{I}_{b_3-2}$. Thus, for the PS-ANOVA model the precision matrix is defined as

$$\boldsymbol{G}_{\mathrm{S}}^{-1} = \mathrm{blockdiag}\left(\frac{1}{\sigma_1^2}\boldsymbol{\Lambda}_1, \frac{1}{\sigma_2^2}\boldsymbol{\Lambda}_2, \frac{1}{\sigma_3^2}\boldsymbol{\Lambda}_3, \frac{1}{\sigma_4^2}\boldsymbol{\Lambda}_4, \frac{1}{\sigma_5^2}\boldsymbol{\Lambda}_5\right) = \sum_{j=1}^{5} \sigma_j^{-2}\tilde{\boldsymbol{\Lambda}}_j,$$

where $\tilde{\boldsymbol{\Lambda}}_j$ is a block diagonal matrix where the jth block is $\boldsymbol{\Lambda}_j$ and the remaining blocks are all-zeroes matrices of proper dimension, e.g., $\tilde{\boldsymbol{\Lambda}}_1 = \mathrm{blockdiag}(\boldsymbol{\Lambda}_1, \boldsymbol{0}, \boldsymbol{0}, \boldsymbol{0}, \boldsymbol{0})$. As a consequence, the variance-covariance matrix is easy to compute (it is a linear function of variance parameters)

$$\boldsymbol{G}_{\mathrm{S}} = \mathrm{blockdiag}\left(\sigma_1^2\boldsymbol{\Sigma}_1, \sigma_2^2\boldsymbol{\Sigma}_2, \sigma_3^2\boldsymbol{\Sigma}_3, \sigma_4^2\boldsymbol{\Sigma}_4, \sigma_5^2\boldsymbol{\Sigma}_5\right), \tag{3.27}$$

where $\boldsymbol{\Sigma}_j = \boldsymbol{\Lambda}_j^{-1}$ $(j = 1, \ldots, 5)$. Note that now $\boldsymbol{G}_{\mathrm{S}}$ has a standard form, and thus standard mixed model software can be used to estimate the PS-ANOVA model.

3.2.3 Mixed model formulation of P-splines in three dimensions

We can extend and combine the ideas of the mixed model formulation from one and two-dimensions to the spatio-temporal three-dimensional function, $\boldsymbol{f}_{\mathrm{ST}}$ (for details, see Lee, 2010; Lee & Durban, 2011)

$$\boldsymbol{f}_{\mathrm{ST}} = \boldsymbol{B}_{\mathrm{ST}}\boldsymbol{\theta}_{\mathrm{ST}} = \boldsymbol{X}_{\mathrm{ST}}\boldsymbol{\beta}_{\mathrm{ST}} + \boldsymbol{Z}_{\mathrm{ST}}\boldsymbol{u}_{\mathrm{ST}}, \quad \boldsymbol{u}_{\mathrm{ST}} \sim N(\boldsymbol{0}, \boldsymbol{G}_{\mathrm{ST}}), \tag{3.28}$$

where the mixed model design matrices $\boldsymbol{X}_{\mathrm{ST}}$ and $\boldsymbol{Z}_{\mathrm{ST}}$ are constructed in a similar fashion to the two dimensional case, i.e., $\boldsymbol{X}_{\mathrm{ST}} = \boldsymbol{B}_{\mathrm{ST}}\boldsymbol{U}_{\mathrm{ST0}}$ and $\boldsymbol{Z}_{\mathrm{ST}} = \boldsymbol{B}_{\mathrm{ST}}\boldsymbol{U}_{\mathrm{ST+}}$, with spatio-temporal cubic B-spline design matrix $\boldsymbol{B}_{\mathrm{ST}} = \boldsymbol{B}_{\mathrm{S}} \otimes \boldsymbol{B}_{\mathrm{T}} = (\boldsymbol{B}_2 \square \boldsymbol{B}_3) \otimes \boldsymbol{B}_{\mathrm{T}}$ as defined in (3.14), and spatio-temporal transformation matrix $\boldsymbol{U}_{\mathrm{ST}} = \boldsymbol{U}_2 \otimes \boldsymbol{U}_3 \otimes \boldsymbol{U}_1$, with $\boldsymbol{U}_2$ and $\boldsymbol{U}_3$ as defined in (3.21), and $\boldsymbol{U}_1 = [\boldsymbol{U}_{10} | \boldsymbol{U}_{1+}]$, such that

$$\begin{aligned}
\boldsymbol{U}_{\mathrm{ST0}} &= [\boldsymbol{U}_{20} \otimes \boldsymbol{U}_{30} \otimes \boldsymbol{U}_{10}] \\
\boldsymbol{U}_{\mathrm{ST+}} &= [\boldsymbol{U}_{20} \otimes \boldsymbol{U}_{30} \otimes \boldsymbol{U}_{1+} \mid \boldsymbol{U}_{20} \otimes \boldsymbol{U}_{3+} \otimes \boldsymbol{U}_{10} \mid \boldsymbol{U}_{20} \otimes \boldsymbol{U}_{3+} \otimes \boldsymbol{U}_{1+} \mid \\
&\qquad \boldsymbol{U}_{2+} \otimes \boldsymbol{U}_{30} \otimes \boldsymbol{U}_{10} \mid \boldsymbol{U}_{2+} \otimes \boldsymbol{U}_{30} \otimes \boldsymbol{U}_{1+} \mid \boldsymbol{U}_{2+} \otimes \boldsymbol{U}_{3+} \otimes \boldsymbol{U}_{10} \mid \boldsymbol{U}_{2+} \otimes \boldsymbol{U}_{3+} \otimes \boldsymbol{U}_{1+}],
\end{aligned}$$

and

$$\begin{aligned}
\boldsymbol{X}_{\mathrm{ST}} &= [(\boldsymbol{X}_2 \square \boldsymbol{X}_3) \otimes \boldsymbol{X}_{\mathrm{T}}] \\
\boldsymbol{Z}_{\mathrm{ST}} &= [(\boldsymbol{Z}_2 \square \boldsymbol{X}_3) \otimes \boldsymbol{X}_{\mathrm{T}} \mid (\boldsymbol{X}_2 \square \boldsymbol{Z}_3) \otimes \boldsymbol{X}_{\mathrm{T}} \mid (\boldsymbol{X}_2 \square \boldsymbol{X}_3) \otimes \boldsymbol{Z}_{\mathrm{T}} \mid \\
&\qquad (\boldsymbol{Z}_2 \square \boldsymbol{Z}_3) \otimes \boldsymbol{X}_{\mathrm{T}} \mid (\boldsymbol{Z}_2 \square \boldsymbol{X}_3) \otimes \boldsymbol{Z}_{\mathrm{T}} \mid (\boldsymbol{X}_2 \square \boldsymbol{Z}_3) \otimes \boldsymbol{Z}_{\mathrm{T}} \mid (\boldsymbol{Z}_2 \square \boldsymbol{Z}_3) \otimes \boldsymbol{Z}_{\mathrm{T}}],
\end{aligned} \tag{3.29}$$

where $\boldsymbol{X}_{\mathrm{T}}$ and $\boldsymbol{Z}_{\mathrm{T}}$ were defined previously in (3.18) for the one-dimensional case, and $\boldsymbol{X}_2$, $\boldsymbol{Z}_2$, $\boldsymbol{X}_3$, $\boldsymbol{Z}_3$ in (3.22) for the bi-dimensional case. Moreover, by replacing in (3.29) the fixed design matrices by, respectively, $\boldsymbol{X}_2 = [\boldsymbol{1}_M \mid \boldsymbol{r}]$, $\boldsymbol{X}_3 = [\boldsymbol{1}_M \mid \boldsymbol{c}]$ and $\boldsymbol{X}_{\mathrm{T}} = [\boldsymbol{1}_n \mid \boldsymbol{t}]$ (or more precisely, by the fixed matrices obtained using Wood et al. (2013) approach), and after some reorganisation of the matrices, we arrive at the following expressions (and the ANOVA-type decomposition proposed by Lee & Durban, 2011)

$$\begin{aligned}
\boldsymbol{X}_{\mathrm{ST}} \equiv [& \underbrace{\boldsymbol{1}_{Mn} \mid \boldsymbol{r} \otimes \boldsymbol{1}_n \mid \boldsymbol{c} \otimes \boldsymbol{1}_n \mid \boldsymbol{t} \otimes \boldsymbol{1}_M \mid (\boldsymbol{r} \odot \boldsymbol{c}) \otimes \boldsymbol{1}_n \mid \boldsymbol{r} \otimes \boldsymbol{t} \mid \boldsymbol{c} \otimes \boldsymbol{t} \mid (\boldsymbol{r} \odot \boldsymbol{c}) \otimes \boldsymbol{t}}_{\text{Linear effects and interactions}}], \\
\boldsymbol{Z}_{\mathrm{ST}} \equiv [& \underbrace{\boldsymbol{Z}_2 \otimes \boldsymbol{1}_n \mid (\boldsymbol{Z}_2 \square \boldsymbol{c}) \otimes \boldsymbol{1}_n \mid \boldsymbol{Z}_2 \otimes \boldsymbol{t} \mid (\boldsymbol{Z}_2 \square \boldsymbol{c}) \otimes \boldsymbol{t}}_{\text{smooth row-related effects}} \mid \underbrace{\boldsymbol{Z}_3 \otimes \boldsymbol{1}_n \mid (\boldsymbol{r} \square \boldsymbol{Z}_3) \otimes \boldsymbol{1}_n \mid \boldsymbol{Z}_3 \otimes \boldsymbol{t} \mid (\boldsymbol{r} \square \boldsymbol{Z}_3) \otimes \boldsymbol{t}}_{\text{smooth column-related effects}} \mid \\
& \underbrace{\boldsymbol{Z}_{\mathrm{T}} \otimes \boldsymbol{1}_M \mid \boldsymbol{r} \otimes \boldsymbol{Z}_{\mathrm{T}} \mid \boldsymbol{c} \otimes \boldsymbol{Z}_{\mathrm{T}} \mid (\boldsymbol{r} \odot \boldsymbol{c}) \otimes \boldsymbol{Z}_{\mathrm{T}}}_{\text{smooth time-related effects}} \mid \underbrace{(\boldsymbol{Z}_2 \square \boldsymbol{Z}_3) \otimes \boldsymbol{1}_M \mid (\boldsymbol{Z}_2 \square \boldsymbol{Z}_3) \otimes \boldsymbol{t}}_{\text{smooth row and column interactions}} \mid \\
& \underbrace{\boldsymbol{Z}_2 \otimes \boldsymbol{Z}_{\mathrm{T}} \mid (\boldsymbol{Z}_2 \square \boldsymbol{c}) \otimes \boldsymbol{Z}_{\mathrm{T}}}_{\text{smooth row and time interactions}} \mid \underbrace{\boldsymbol{Z}_3 \otimes \boldsymbol{Z}_{\mathrm{T}} \mid (\boldsymbol{r} \square \boldsymbol{Z}_3) \otimes \boldsymbol{Z}_{\mathrm{T}}}_{\text{smooth column and time interactions}} \mid \underbrace{(\boldsymbol{Z}_2 \square \boldsymbol{Z}_3) \otimes \boldsymbol{Z}_{\mathrm{T}}}_{\text{smooth space-time interactions}}].
\end{aligned} \tag{3.30}$$

Regarding the vectors of fixed and random effect coefficients, we have

$$\begin{aligned}
\boldsymbol{\beta}_{\text{ST}} &= (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7)^T \,, \\
\boldsymbol{u}_{\text{ST}} &= \Bigg(\underbrace{\boldsymbol{u}_2^T, \boldsymbol{u}_{2;3}^T, \boldsymbol{u}_{2;1}^T, \boldsymbol{u}_{2;31}^T}_{\boldsymbol{u}_r}, \underbrace{\boldsymbol{u}_3^T, \boldsymbol{u}_{3;2}^T, \boldsymbol{u}_{3;1}^T, \boldsymbol{u}_{3;21}^T}_{\boldsymbol{u}_c}, \underbrace{\boldsymbol{u}_1^T, \boldsymbol{u}_{1;2}^T, \boldsymbol{u}_{1;3}^T, \boldsymbol{u}_{1;2,3}^T}_{\boldsymbol{u}_t} \\
&\quad \underbrace{\boldsymbol{u}_{2|3}^T, \boldsymbol{u}_{2|3;1}^T}_{\boldsymbol{u}_{r|c}}, \underbrace{\boldsymbol{u}_{2|1}^T, \boldsymbol{u}_{2|1;3}^T}_{\boldsymbol{u}_{r|t}}, \underbrace{\boldsymbol{u}_{3|1}^T, \boldsymbol{u}_{3|1,2}^T}_{\boldsymbol{u}_{c|t}}, \underbrace{\boldsymbol{u}_{2|3|1}^T}_{\boldsymbol{u}_{r|c|t}} \Bigg)^T .
\end{aligned} \tag{3.31}$$

We note that for the vector of random effects, $\boldsymbol{u}_{\text{ST}}$, we have 19 sets of random effects, each associated with one block in (3.30). These sets can be further grouped into 7 larger sets. The first three, $\boldsymbol{u}_r$, $\boldsymbol{u}_c$ and $\boldsymbol{u}_t$, correspond to one-dimensional smooth (non-linear) effects along the rows, columns and time, respectively, and $\boldsymbol{u}_{r|c}$, $\boldsymbol{u}_{r|t}$ and $\boldsymbol{u}_{c|t}$ to bivariate smooth (non-linear) interactions between rows and columns, rows and time, and columns and time, respectively. Finally, $\boldsymbol{u}_{r|c|t}$ corresponds to the trivariate smooth (non-linear) interaction between rows, columns and time. With these 7 sets in mind, the precision matrix (i.e., the inverse of variance-covariance matrix) associated with $\boldsymbol{u}_{\text{ST}}$, is a block diagonal matrix, with 7 blocks, each related to each set of random effects

$$\begin{aligned}
\boldsymbol{G}_{\text{ST}}^{-1} =& \text{blockdiag} \Bigg(\frac{1}{\sigma^2_{\text{ST},2}} \boldsymbol{\Lambda}_{2+} \otimes \boldsymbol{I}_2 \otimes \boldsymbol{I}_2, \frac{1}{\sigma^2_{\text{ST},3}} \boldsymbol{I}_2 \otimes \boldsymbol{\Lambda}_{3+} \otimes \boldsymbol{I}_2, \frac{1}{\sigma^2_{\text{ST},1}} \boldsymbol{I}_2 \otimes \boldsymbol{I}_2 \otimes \boldsymbol{\Lambda}_{1+}, \\
& \frac{1}{\sigma^2_{\text{ST},2}} \boldsymbol{\Lambda}_{2+} \otimes \boldsymbol{I}_{b_3-2} \otimes \boldsymbol{I}_2 + \frac{1}{\sigma^2_{\text{ST},3}} \boldsymbol{I}_{b_2-2} \otimes \boldsymbol{\Lambda}_{3+} \otimes \boldsymbol{I}_2, \\
& \frac{1}{\sigma^2_{\text{ST},2}} \boldsymbol{\Lambda}_{2+} \otimes \boldsymbol{I}_2 \otimes \boldsymbol{I}_{b_1-2} + \frac{1}{\sigma^2_{\text{ST},1}} \boldsymbol{I}_{b_2-2} \otimes \boldsymbol{I}_2 \otimes \boldsymbol{\Lambda}_{1+}, \\
& \frac{1}{\sigma^2_{\text{ST},3}} \boldsymbol{I}_2 \otimes \boldsymbol{\Lambda}_{3+} \otimes \boldsymbol{I}_{b_1-2} + \frac{1}{\sigma^2_{\text{ST},1}} \boldsymbol{I}_2 \otimes \boldsymbol{I}_{b_3-2} \otimes \boldsymbol{\Lambda}_{1+}, \\
& \frac{1}{\sigma^2_{\text{ST},2}} \boldsymbol{\Lambda}_{2+} \otimes \boldsymbol{I}_{b_3-2} \otimes \boldsymbol{I}_{b_1-2} + \frac{1}{\sigma^2_{\text{ST},3}} \boldsymbol{I}_{b_2-2} \otimes \boldsymbol{\Lambda}_{3+} \otimes \boldsymbol{I}_{b_1-2} + \frac{1}{\sigma^2_{\text{ST},1}} \boldsymbol{I}_{b_2-2} \otimes \boldsymbol{I}_{b_3-2} \otimes \boldsymbol{\Lambda}_{1+} \Bigg) \\
=& \sum_{j \in \{2,3,1\}} \sigma^{-2}_{\text{ST},j} \tilde{\boldsymbol{\Lambda}}_{j+},
\end{aligned} \tag{3.32}$$

where $\boldsymbol{G}_{\text{ST}}^{-1}$ depends on three precision parameters, $\sigma^{-2}_{\text{ST},2} = \lambda_2/\sigma^2$, $\sigma^{-2}_{\text{ST},3} = \lambda_3/\sigma^2$, and $\sigma^{-2}_{\text{ST},1} = \lambda_1/\sigma^2$ (responsible for controlling the smoothness along the rows, columns and time, respectively), and

$$\begin{aligned}
\tilde{\boldsymbol{\Lambda}}_{2+} =& \text{blockdiag} \left(\boldsymbol{\Lambda}_{2+} \otimes \boldsymbol{I}_2 \otimes \boldsymbol{I}_2, \boldsymbol{0}, \boldsymbol{0}, \; \boldsymbol{\Lambda}_{2+} \otimes \boldsymbol{I}_{b_3-2} \otimes \boldsymbol{I}_2, \boldsymbol{\Lambda}_{2+} \otimes \boldsymbol{I}_2 \otimes \boldsymbol{I}_{b_1-2}, \boldsymbol{0}, \; \boldsymbol{\Lambda}_{2+} \otimes \boldsymbol{I}_{b_3-2} \otimes \boldsymbol{I}_{b_1-2} \right) \\
\tilde{\boldsymbol{\Lambda}}_{3+} =& \text{blockdiag} \left(\boldsymbol{0}, \boldsymbol{I}_2 \otimes \boldsymbol{\Lambda}_{3+} \otimes \boldsymbol{I}_2, \boldsymbol{0}, \; \boldsymbol{I}_{b_2-2} \otimes \boldsymbol{\Lambda}_{3+} \otimes \boldsymbol{I}_2, \boldsymbol{0}, \boldsymbol{I}_2 \otimes \boldsymbol{\Lambda}_{3+} \otimes \boldsymbol{I}_{b_1-2}, \; \boldsymbol{I}_{b_2-2} \otimes \boldsymbol{\Lambda}_{3+} \otimes \boldsymbol{I}_{b_1-2} \right) \\
\tilde{\boldsymbol{\Lambda}}_{1+} =& \text{blockdiag} \left(\boldsymbol{0}, \boldsymbol{0}, \boldsymbol{I}_2 \otimes \boldsymbol{I}_2 \otimes \boldsymbol{\Lambda}_{1+}, \; \boldsymbol{0}, \boldsymbol{I}_{b_2-2} \otimes \boldsymbol{I}_2 \otimes \boldsymbol{\Lambda}_{1+}, \boldsymbol{I}_2 \otimes \boldsymbol{I}_{b_3-2} \otimes \boldsymbol{\Lambda}_{1+}, \; \boldsymbol{I}_{b_2-2} \otimes \boldsymbol{I}_{b_3-2} \otimes \boldsymbol{\Lambda}_{1+} \right) .
\end{aligned} \tag{3.33}$$

It is worth indicating that $\boldsymbol{f}_{\mathrm{ST}}$ also accepts an ANOVA-type decomposition similar to the one shown for the two-dimensional case

$$
\begin{aligned}
\boldsymbol{f}_{\mathrm{ST}} = \underbrace{\mathbf{1}_{Mn}\beta_0 + (\boldsymbol{r}\otimes\mathbf{1}_n)\beta_1 + (\boldsymbol{c}\otimes\mathbf{1}_n)\beta_2 + (\boldsymbol{t}\otimes\mathbf{1}_M)\beta_3 + ((\boldsymbol{r}\odot\boldsymbol{c})\otimes\mathbf{1}_n)\beta_4 + (\boldsymbol{r}\otimes\boldsymbol{t})\beta_5 + (\boldsymbol{c}\otimes\boldsymbol{t})\beta_6 + ((\boldsymbol{r}\odot\boldsymbol{c})\otimes\boldsymbol{t})\beta_7}_{\text{Linear effects and interactions}} + \\
\underbrace{\underbrace{f_2(\boldsymbol{r})\otimes\mathbf{1}_n}_{(\mathbf{Z}_2\otimes\mathbf{1}_n)\boldsymbol{u}_2} + \underbrace{(\boldsymbol{c}\odot h_{2;3}(\boldsymbol{r}))\otimes\mathbf{1}_n}_{((\mathbf{Z}_2\square\boldsymbol{c})\otimes\mathbf{1}_n)\boldsymbol{u}_{2;3}} + \underbrace{h_{2;1}(\boldsymbol{r})\otimes\boldsymbol{t}}_{(\mathbf{Z}_2\otimes\boldsymbol{t})\boldsymbol{u}_{2;1}} + \underbrace{(\boldsymbol{c}\odot h_{2;31}(\boldsymbol{r}))\otimes\boldsymbol{t}}_{((\mathbf{Z}_2\square\boldsymbol{c})\otimes\boldsymbol{t})\boldsymbol{u}_{2;31}}}_{\text{Non-linear row-related effects}} + \\
\underbrace{\underbrace{f_3(\boldsymbol{c})\otimes\mathbf{1}_n}_{(\mathbf{Z}_3\otimes\mathbf{1}_n)\boldsymbol{u}_3} + \underbrace{(\boldsymbol{r}\odot h_{3;2}(\boldsymbol{c}))\otimes\mathbf{1}_n}_{((\boldsymbol{r}\square\mathbf{Z}_3)\otimes\mathbf{1}_n)\boldsymbol{u}_{3;2}} + \underbrace{h_{3;1}(\boldsymbol{c})\otimes\boldsymbol{t}}_{(\mathbf{Z}_3\otimes\boldsymbol{t})\boldsymbol{u}_{3;1}} + \underbrace{(\boldsymbol{r}\odot h_{3;21}(\boldsymbol{c}))\otimes\boldsymbol{t}}_{((\boldsymbol{r}\square\mathbf{Z}_3)\otimes\boldsymbol{t})\boldsymbol{u}_{3;21}}}_{\text{Non-linear column-related effects}} + \\
\underbrace{\underbrace{f_1(\boldsymbol{t})\otimes\mathbf{1}_M}_{(\mathbf{Z}_\mathrm{T}\otimes\mathbf{1}_M)\boldsymbol{u}_1} + \underbrace{\boldsymbol{r}\otimes h_{1;2}(\boldsymbol{t})}_{(\boldsymbol{r}\otimes\mathbf{Z}_\mathrm{T})\boldsymbol{u}_{1;2}} + \underbrace{\boldsymbol{c}\otimes h_{1;3}(\boldsymbol{t})}_{(\boldsymbol{c}\otimes\mathbf{Z}_\mathrm{T})\boldsymbol{u}_{1;3}} + \underbrace{(\boldsymbol{r}\odot\boldsymbol{c})\otimes h_{1;23}(\boldsymbol{t})}_{((\boldsymbol{r}\odot\boldsymbol{c})\otimes\mathbf{Z}_\mathrm{T})\boldsymbol{u}_{1;23}}}_{\text{Non-linear time-related effects}} + \\
\underbrace{\underbrace{h_{2|3}(\boldsymbol{r},\boldsymbol{c})\otimes\mathbf{1}_M}_{((\mathbf{Z}_2\square\mathbf{Z}_3)\otimes\mathbf{1}_M)\boldsymbol{u}_{2|3}} + \underbrace{h_{2|3;1}(\boldsymbol{r},\boldsymbol{c})\otimes\boldsymbol{t}}_{((\mathbf{Z}_2\square\mathbf{Z}_3)\otimes\boldsymbol{t})\boldsymbol{u}_{2|3;1}}}_{\text{Non-linear row and column interactions}} + \underbrace{\underbrace{h_{2|1}(\boldsymbol{r},\boldsymbol{t})}_{(\mathbf{Z}_2\otimes\mathbf{Z}_\mathrm{T})\boldsymbol{u}_{2|1}} + \underbrace{\boldsymbol{c}\odot f_{2|1;3}(\boldsymbol{r},\boldsymbol{t})}_{((\mathbf{Z}_2\square\boldsymbol{c})\otimes\mathbf{Z}_\mathrm{T})\boldsymbol{u}_{2|1;3}}}_{\text{Non-linear row and time interactions}} + \\
\underbrace{\underbrace{h_{3|1}(\boldsymbol{c},\boldsymbol{t})}_{(\mathbf{Z}_3\otimes\mathbf{Z}_\mathrm{T})\boldsymbol{u}_{3|1}} + \underbrace{\boldsymbol{r}\odot f_{3|1;2}(\boldsymbol{c},\boldsymbol{t})}_{((\boldsymbol{r}\square\mathbf{Z}_3)\otimes\mathbf{Z}_\mathrm{T})\boldsymbol{u}_{3|1;2}}}_{\text{Non-linear column and time interactions}} + \underbrace{\underbrace{h_{2|3|1}(\boldsymbol{r},\boldsymbol{c},\boldsymbol{t})}_{((\mathbf{Z}_2\square\mathbf{Z}_3)\otimes\mathbf{Z}_\mathrm{T})\boldsymbol{u}_{2|3|1}}}_{\text{Non-linear space-time interaction}} .
\end{aligned}
\tag{3.34}
$$

In contrast to the two-dimensional function $\boldsymbol{f}_{\mathrm{S}}$ in (3.26), more components are now present (there are three variables involved: rows, columns and time), yet their interpretation is similar. For instance, $c \times h_{2;31}(r) \times t$ are linear interaction trends in the columns (c) and time (t), but with slopes ($h_{2;31}(r)$) that change smoothly along the rows.

3.3 Coefficients and variance parameters estimation

Estimation of the standard linear mixed model (3.17) can be carried out with any mixed-model software, previosuly mentioned in Section 1.2.2. Also, we can use the R-package `ASReml-R` (Butler et al., 2018) for that purpose. However, the P-spline context requires additional estimation methods, especially for multidimensional cases (which are our main target). For instance, in some situations, standard mixed model procedures can not be used due to the variance-covariance matrices having a non-standard form (e.g., $\boldsymbol{G}_{\mathrm{S}}$ in (3.25) and $\boldsymbol{G}_{\mathrm{ST}}$ in (3.32) have a block involving several variance parameters but in a non-linear way). Consequently, in the context of multidimensional P-splines, it is common to work with the precision matrix, $\boldsymbol{G}^{-1}$, which is linear in the precision parameters. In the general setting of mixed models, estimation is based

on the variance-covariance matrix for the random effects, $\boldsymbol{G}$, which is linear in the variance parameter (e.g. the PS-ANOVA specification results in a standard form for the variance-covariance matrix $\boldsymbol{G}_{\mathrm{S}}$ in (3.27)).

To tackle this particularity, we use two proposed algorithms: SAP (Separation of Anisotropic Penalties; Rodríguez-Álvarez et al., 2015), and SOP (Separation of Overlapping Precision Matrices; Rodríguez-Álvarez et al., 2019). In fact, SOP algorithm generalises SAP, and the SOP method reduces to the estimating algorithm described in Harville (1977) when no overlapping precision matrices are present (which is the case of PS-ANOVA). Then, for clarity and simplicity, hereafter, we will use SOP to refer to both SOP and SAP, and we will use the SOP algorithm to present the general proposal. In what follows, we present the estimation algorithm, but we start rewriting the general multidimensional mixed model as follows

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\varepsilon}; \quad \boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \boldsymbol{R}), \quad \boldsymbol{u} \sim N(\boldsymbol{0}, \boldsymbol{G}),$$

where $\boldsymbol{R} = \sigma^2\boldsymbol{I}$, and $\boldsymbol{G}^{-1} = \sum_j \sigma_j^{-2}\tilde{\boldsymbol{\Lambda}}_{j+}$. For instance the precision matrices $\boldsymbol{G}_{\mathrm{S}}^{-1}$ in (3.23) with $j \in \{2, 3\}$, and $\boldsymbol{G}_{\mathrm{ST}}^{-1}$ in (3.32) with $j \in \{1, 2, 3\}$.

3.3.1 Estimating algorithm

Initialise Set initial values for the variance parameters $\hat{\sigma}_j^{2[0]}$ and for the residual variance $\hat{\sigma}^{2[0]}$. Set $it = 0$.

Step 1. Given the initial estimates of variance parameters, estimate the empirical best linear unbiased estimates (BLUEs), $\boldsymbol{\beta}$, and predictors (BLUPs), $\boldsymbol{u}$, by the solution of the Henderson's mixed model equations (Henderson, 1963)

$$\underbrace{\begin{pmatrix} \boldsymbol{X}^T\boldsymbol{R}^{[it]^{-1}}\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{R}^{[it]^{-1}}\boldsymbol{Z} \\ \boldsymbol{Z}^T\boldsymbol{R}^{[it]^{-1}}\boldsymbol{X} & \boldsymbol{Z}^T\boldsymbol{R}^{[it]^{-1}}\boldsymbol{Z} + \boldsymbol{G}^{[it]^{-1}} \end{pmatrix}}_{\boldsymbol{C}^{[it]}} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{u}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}^T\boldsymbol{R}^{[it]^{-1}}\boldsymbol{y} \\ \boldsymbol{Z}^T\boldsymbol{R}^{[it]^{-1}}\boldsymbol{y} \end{bmatrix}. \tag{3.35}$$

Step 2 Update the variance parameters by maximising the restricted maximum likelihood function (REML; Patterson & Thompson, 1971),

$$\hat{\sigma}_j^2 = \frac{\hat{\boldsymbol{u}}^{[it]T}\tilde{\boldsymbol{\Lambda}}_{j+}\hat{\boldsymbol{u}}^{[it]}}{\mathrm{ED}_j^{[it]}},$$

and residual variance,

$$\hat{\sigma}^2 = \frac{\|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^{[it]} - \boldsymbol{Z}\hat{\boldsymbol{u}}^{[it]}\|^2}{\# - \mathrm{rank}(\boldsymbol{X}) - \sum_j \mathrm{ED}_j^{[it]}},$$

where # denotes the number of observations (e.g. for the two-dimensional model # : M, for the three-dimensional model # : nM), and effective dimension

$$\mathrm{ED}_j^{[it]} = \mathrm{trace}\left(\left(\boldsymbol{G}^{[it]} - \boldsymbol{C}^{+[it]^{-1}}\right)\frac{\tilde{\boldsymbol{\Lambda}}_{j+}}{\hat{\sigma}_j^{2[it]}}\right), \tag{3.36}$$

where $\boldsymbol{C}^{[it]^{-1}}$ denotes the inverse of $\boldsymbol{C}^{[it]}$ in (3.35), and $\boldsymbol{C}^{+[it]^{-1}}$ is the partition of $\boldsymbol{C}^{[it]^{-1}}$ corresponding to the random vector of coefficients $\boldsymbol{u}$.

Step 3. Repeat Steps 1 to 2, with variance parameters being replaced for those obtained in the previous iteration until the convergence criterion: difference in the REML deviance between two consecutive iterations is small enough.

3.4 Standard errors and pointwise confidence intervals

We follow the work by Ruppert et al. (2003) and Welham et al. (2004) to obtain approximate $100(1-\alpha)\%$ pointwise confidence intervals and standard errors for predictions (i.e., any quantity that is linear in the model coefficients). In particular, for fixed values of the variance parameters and error variance, we have that

$$\text{Cov}\left(\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ (\hat{\boldsymbol{u}} - \boldsymbol{u}) \end{bmatrix}\right) \approx \boldsymbol{C}^{-1},$$

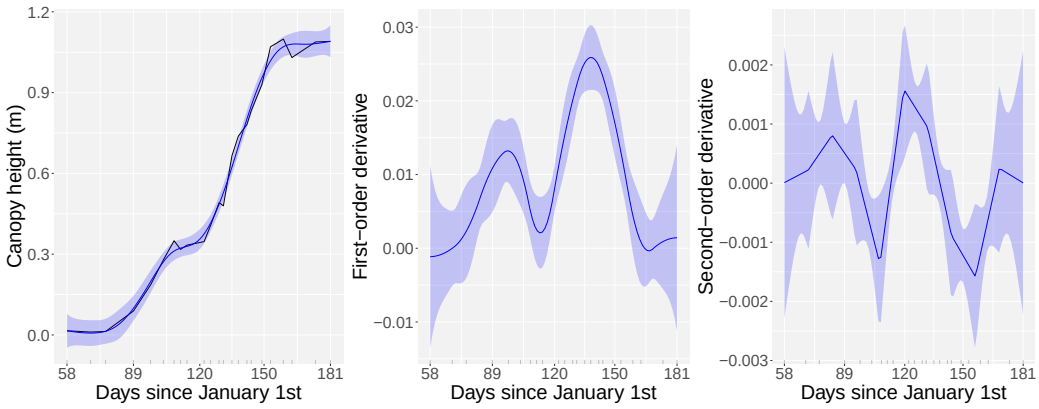
with $\boldsymbol{C}$ as defined in (3.35). We use this result to obtain pointwise confidence intervals for the predictions. As an example, for the one-dimensional function (3.19), suppose we want to obtain a confidence interval for f_{T} at a particular time, t_*. Let $\boldsymbol{X}_{\text{T}(t_*)}$ and $\boldsymbol{Z}_{\text{T}(t_*)}$ denote the mixed model matrices for this particular value, and let $\hat{f}_{\text{T}}(t_*) = \boldsymbol{X}_{\text{T}(t_*)}\hat{\boldsymbol{\beta}}_{\text{T}} + \boldsymbol{Z}_{\text{T}(t_*)}\hat{u}_{\text{T}}$. Then, an approximate $100(1-\alpha)\%$ pointwise confidence interval for $f_{\text{T}}(t_*)$ is

$$\hat{f}_{\text{T}}(t_*) \pm z_{1-\alpha/2}\widehat{\text{SE}}_{\hat{f}_{\text{T}}(t_*)}, \tag{3.37}$$

where $z_{1-\alpha/2}$ denotes the $1-\alpha/2$ quantile of a $N(0,1)$, and

$$\widehat{\text{SE}}_{\hat{f}_{\text{T}}(t_*)} \equiv \sqrt{\widehat{\text{Var}}\left\{\hat{f}_{\text{T}}(t_*) - f_{\text{T}}(t_*)\right\}} = \sqrt{\text{diag}\left([\boldsymbol{X}_{\text{T}(t_*)}|\boldsymbol{Z}_{\text{T}(t_*)}]\boldsymbol{C}^{-1}[\boldsymbol{X}_{\text{T}(t_*)}|\boldsymbol{Z}_{\text{T}(t_*)}]^T\right)}$$

Figure 3.9(a) depicts the 95% pointwise confidence intervals for the predictions for the canopy height curve shown in Figure 3.1(a). In the same way, confidence intervals for the first and second-order derivatives curves ($\hat{f}'_{\text{T}}(t_*)$, and $\hat{f}''_{\text{T}}(t_*)$) can be obtained as in (3.37) by replacing $\boldsymbol{X}_{\text{T}(t_*)}$ and $\boldsymbol{Z}_{\text{T}(t_*)}$ by $\boldsymbol{X}^h_{\text{T}(t_*)}$ and $\boldsymbol{Z}^h_{\text{T}(t_*)}$, as shown in Figure 3.9(b).



(a) Canopy height

(b) First and second-order derivatives

Figure 3.9: For the canopy height curve (continuous black line) shown in Figure 3.1(a), **(a)** predictions (continuous blue line) with 95% pointwise confidence intervals (blue shaded area), **(b)** first and second-order derivatives (continuous blue lines) with 95% pointwise confidence intervals (blue shaded areas).

Chapter 4

Spatio-temporal modelling of high-throughput phenotyping data: Two-stage approach

This chapter presents a two-stage P-spline-based approach for analysing spatio-temporal HTP data. Section 4.1 describes the first stage, in which we correct for design features and spatial trends per time point. To that aim, we use a SpATS model, where the spatial trends are modelled using the (spatial) two-dimensional smooth function, f_S, previously described in Sections 3.1.2.1 and 3.2.2. The second stage is presented in Section 4.2; in this stage, we focus on the longitudinal modelling of the genetic signal on the spatially corrected data. We extend the longitudinal smooth function, f_T (see Sections 3.1.1 and 3.2.1), to the case of M plants, where plants are nested into genotypes and genotypes are nested into populations. Therefore, we propose a flexible hierarchical three-level P-spline-based curve model, thereby taking advantage of shared longitudinal features between genotypes and plants within genotypes. We let the assessment of the performance of this approach through a simulation study for Chapter 6. In Chapter 7, we illustrate its usage by analysing the HTP data described in Chapter 2. Finally, the results showed in this thesis with this approach can be reproduced with the software developments presented in Chapter 8. This chapter hinges on the material in Perez-Valencia et al. (2022).

4.1 First stage: Spatial correction

The aim of the first-stage is to correct for spatial trends and design features (covariates/factors) that we are not interested in modelling in the second stage of the two-stage approach, separately per time point. On the otherhand, we are interested in the temporal evolution of the genetic signal, while keeping the hierarchical data structure. As such, we model the observed phenotypic trait, y_i, for the ith plant at a specific time point (for simplicity, we omit here the dependence on time), by considering the following spatial model

$$y_i = h_{p(i)} + h_{g(i)} + h_{r(i)} + h_{c(i)} + \underbrace{h_{\mathrm{S}}(r(i), c(i))}_{\text{Spatial trend}} + \sum_e h_{e(i)} + \varepsilon_i,\ \varepsilon_i \sim N(0, \sigma_t^2),\ 1 \leq i \leq M, \tag{4.1}$$

where h_p is the fixed effect coefficient for population p, h_g is the fixed/random effect coefficient (we discuss this choice later in Section 4.1.2) for genotype g, h_r and h_c are random effect coefficients for row r and column c, respectively ($h_r \sim N(0, \sigma^2_{\text{row}})$ and $h_c \sim N(0, \sigma^2_{\text{col}})$; they are included to account for design factors), and $h_{\mathrm{S}}(r, c)$, is a two-dimensional smooth function, defined over the row and column positions, that simultaneously accounts for the spatial (local and global) trend variation across both directions. Other experimental design factors (e.g., presence of block and/or replication and/or lot effects) can be included in the model ($\sum_e h_e$) to build more complex models.

4.1.1 Spatial P-spline-based model: SpATS model

For the first stage, we specifically use a SpATS model to fit the spatial model in (4.1) on $\boldsymbol{y}_t = (y_1(t), \ldots, y_M(t))^T$ separately for each measurement time $t \in \{t_1, \ldots, t_n\}$. A SpATS model is a linear mixed model that is based on the two-dimensional P-spline smooth surface $h_{\mathrm{S}}(r, c)$ at time t, presented in Section 3.1.2.1. In its more general specification and considering genotypes as random, the mixed model formulation of the SpATS model has the following form (for more details, see Rodríguez-Álvarez et al., 2018)

$$\boldsymbol{y}_t = \mathbf{1}_M \beta_{0t} + \underbrace{\boldsymbol{X}_{\mathrm{S}}\boldsymbol{\beta}_{\mathrm{S},t} + \boldsymbol{Z}_{\mathrm{S}}\boldsymbol{u}_{\mathrm{S},t}}_{\text{Spatial trend, } \boldsymbol{h}_{\mathrm{S},t}} + \boldsymbol{X}_{\text{pop}}\boldsymbol{\beta}_{\text{pop},t} + \boldsymbol{Z}_{\text{gen}}\boldsymbol{u}_{\text{gen},t} + \boldsymbol{Z}_{\text{row}}\boldsymbol{u}_{\text{row},t} + \boldsymbol{Z}_{\text{col}}\boldsymbol{u}_{\text{col},t} + \boldsymbol{X}_{\hat{\mathrm{S}}}\boldsymbol{\beta}_{\hat{\mathrm{S}},t} + \boldsymbol{Z}_{\hat{\mathrm{S}}}\boldsymbol{u}_{\hat{\mathrm{S}},t} + \boldsymbol{\varepsilon}_t, \tag{4.2}$$

where $\boldsymbol{X}_{\mathrm{S}}\boldsymbol{\beta}_{\mathrm{S},t}$ (excluding the intercept), $\boldsymbol{Z}_{\mathrm{S}}\boldsymbol{u}_{\mathrm{S},t}$ (with $\boldsymbol{u}_{\mathrm{S},t} \sim N(\mathbf{0}, \boldsymbol{G}_{\mathrm{S},t})$), and $\boldsymbol{G}_{\mathrm{S},t}$ were defined in (3.24) and (3.27). Here, $[\boldsymbol{X}_{\text{pop}} \mid \boldsymbol{X}_{\hat{\mathrm{S}}}]$, with $\boldsymbol{X}_{\hat{\mathrm{S}}} = [\boldsymbol{X}_1 \mid \ldots \mid \boldsymbol{X}_{E_1}]$, are $1 + E_1$ design matrices associated with the population effects and other experimental design factors (categorical covariates), with fixed effect coefficients $(\boldsymbol{\beta}^T_{\text{pop},t}, \boldsymbol{\beta}^T_{\hat{\mathrm{S}},t})^T$ (and $\boldsymbol{\beta}_{\hat{\mathrm{S}},t} = (\boldsymbol{\beta}^T_{1,t}, \ldots, \boldsymbol{\beta}^T_{E_1,t})^T$). The length of each vector of fixed effect coefficients (and therefore the number of columns in the associated design matrices) corresponds to the number of different categories, say $\hat{c}_{e_1}$ ($e_1 = 1, \ldots, E_1$) and $\hat{c}_{\text{pop}}$, of each (fixed) experimental design factor minus one (as the intercept is included in the model). Regarding the random effects,

$[\boldsymbol{Z}_{\text{gen}} \mid \boldsymbol{Z}_{\text{row}} \mid \boldsymbol{Z}_{\text{col}} \mid \boldsymbol{Z}_{\hat{\S}}]$, with $\boldsymbol{Z}_{\hat{\S}} = [\boldsymbol{Z}_1 \mid \ldots \mid \boldsymbol{Z}_{E_2}]$, are design matrices assigning plants to genotypes, rows and columns positions, as well as to $e_2 = 1, \ldots, E_2$ covariates, with random effect coefficients $(\boldsymbol{u}_{\text{gen},t}^T, \boldsymbol{u}_{\text{row},t}^T, \boldsymbol{u}_{\text{col},t}^T, \boldsymbol{u}_{\hat{\S},t}^T)^T \sim N(\boldsymbol{0}, \boldsymbol{G}_{*t})$, with $\boldsymbol{G}_{*t} = \text{blockdiag}(\boldsymbol{G}_{\text{gen},t}, \boldsymbol{G}_{\text{row},t}, \boldsymbol{G}_{\text{col},t}, \boldsymbol{G}_{\hat{\S},t})$, where $\boldsymbol{u}_{\hat{\S},t} = (\boldsymbol{u}_{1,t}^T, \ldots, \boldsymbol{u}_{E_2,t}^T)^T$ and $\boldsymbol{G}_{\hat{\S},t} = \text{blockdiag}(\boldsymbol{G}_{1,t}, \ldots, \boldsymbol{G}_{E_2,t})$, and $\boldsymbol{\varepsilon}_t \sim N(\boldsymbol{0}, \sigma_t^2 \boldsymbol{I}_M)$. Moreover, if genotypes are grouped by populations, different genetic variances are assumed for each group/population, that is $\boldsymbol{G}_{\text{gen},t} = \text{blockdiag}(\sigma_{1,t}^2 \boldsymbol{I}_{\ell_1}, \ldots, \sigma_{K,t}^2 \boldsymbol{I}_{\ell_K})$. Thus, the compound variance-covariance matrix for the full model (4.2) is

$$\boldsymbol{G}_t = \text{blockdiag}(\boldsymbol{G}_{\text{S},t}, \boldsymbol{G}_{*t}) = \sum_{j=1}^{5+K+2+E_2} \sigma_j^2 \tilde{\boldsymbol{\Sigma}}_j, \tag{4.3}$$

where $\boldsymbol{G}_{\text{S},t}$ has 5 variance parameters (see equation (3.27)), $\boldsymbol{G}_{\text{gen},t} = \text{blockdiag}(\sigma_{1,t}^2 \boldsymbol{I}_{\ell_1}, \ldots, \sigma_{K,t}^2 \boldsymbol{I}_{\ell_K}) = \text{blockdiag}(\sigma_{6,t}^2 \boldsymbol{\Sigma}_6, \ldots, \sigma_{(5+K),t}^2 \boldsymbol{\Sigma}_{5+K})$, $\boldsymbol{G}_{\text{row},t} = \sigma_{\text{row},t}^2 \boldsymbol{I}_R = \sigma_{(5+K+1),t}^2 \boldsymbol{\Sigma}_{5+K+1}$, $\boldsymbol{G}_{\text{col},t} = \sigma_{\text{col},t}^2 \boldsymbol{I}_C = \sigma_{(5+K+2),t}^2 \boldsymbol{\Sigma}_{5+K+2}$, $\boldsymbol{G}_{\hat{\S},t} = \text{blockdiag}(\sigma_{1,t}^2 \boldsymbol{I}_{\hat{c}_1}, \ldots, \sigma_{E_2,t}^2 \boldsymbol{I}_{\hat{c}_{E_2}}) = \text{blockdiag}(\sigma_{(5+K+2+1),t}^2 \boldsymbol{\Sigma}_{5+K+2+1}, \ldots, \sigma_{(5+K+2+E_2),t}^2 \boldsymbol{\Sigma}_{5+K+2+E_2})$, and $\tilde{\boldsymbol{\Sigma}}_j$ is a block diagonal matrix where the jth block is $\boldsymbol{\Sigma}_j$ and the remaining blocks are all-zeroes matrices of proper dimensions.

In plant breeding, it is usual to model genotypes as fixed effect coefficients in the first stage of a two stage approach, where the first stage consists of a per trial analysis and the second stage consists of a weighted across trial analysis. In the second stage genotypes are then taken as random effect coefficients. The population (if any) as well as the replicate and/or (incomplete) block effects are also modelled as fixed effect coefficients. The column and row design factors are commonly modelled as random effects. We propose to model the genotype effects as random. In the following section some implications of this choice are discussed, and a comparison with both results (genotypes as fixed and random) are shown in Chapter 7.

4.1.2 Genotypes as random or fixed effect coefficients

One important question that arises at this first stage is whether to model genotypes as fixed effec coefficient (as usually in stage-wise analyses, see e.g., Damesa et al., 2017; Roth et al., 2021; van Eeuwijk et al., 2019) or random effect coefficient (as we finally decided to do). Let us address the two models.

We start by considering genotypes as fixed effect coefficient. For the general SpATS model (4.2), let $\boldsymbol{X}_{\text{gen}}\boldsymbol{\beta}_{\text{gen},t}$ replace the $\boldsymbol{Z}_{\text{gen}}\boldsymbol{u}_{\text{gen}}$ term, where $\boldsymbol{X}_{\text{gen}}$ is the design matrix (of dimension $M \times (L-1)$; the intercept is included in the model) assigning plants to genotypes, and $\boldsymbol{\beta}_{\text{gen},t} = (\beta_{1,t}, \ldots, \beta_{(L-1),t})^T$ is the vector of genotypic fixed effect coefficient. We note that if other fixed effect are included in the model, the intercept will represent one of the levels for each of these factors. When genotypes are modelled as random effects, $\boldsymbol{Z}_{\text{gen}}\boldsymbol{u}_{\text{gen},t}$ is one of the terms in the SpATS model (4.2). Here, $\boldsymbol{Z}_{\text{gen}}$ is the design matrix (of dimension $M \times L$) assigning plants to genotypes, and $\boldsymbol{u}_{\text{gen},t} = (u_{1,t}, \ldots, u_{L,t})^T \sim N(\boldsymbol{0}, \boldsymbol{G}_{\text{gen},t})$ is the vector genotypic random effects.

The reason to model genotypes as fixed effect coefficients in stage-wise analyses is the "double-shrinkage" (for two-stage approaches) of genotype effects when these are considered random (Damesa et al., 2017; Piepho et al., 2012; A. Smith et al., 2001). However, most of these analyses keep the data resolution at the genotype level (i.e., genotype means are computed/predicted) for the first stage. Instead, we propose to correct the phenotype of interest in such a way that the data resolution is kept at the plant level by the inclusion of the residual component to the genotype prediction (as will be explained later in Section 4.1.3). In our experience, comparison of the spatially corrected phenotype when modelling genotypes as fixed or random effect coefficients shows essentially identical results (as will be illustrated later in Chapter 7). We believe this is because the shrinkage of the genotypic BLUPs is counteracted by the inclusion of the residual component into the correction. The implementation of the first stage allows for both modelling strategies: fixed and random genotypic effect coefficients. However, our final results are consistently reported by considering genotypes as random at this first stage. Besides the fact that BLUPs often improve precision compared to BLUEs (as shown by Piepho et al., 2008, in plant breeding applications), incorporating a different genetic variance per population (if any) will allow for more flexibility/fidelity to data behaviour. Finally, BLUPs allow heritabilities to be computed for each measurement time. Heritability is a measure of the proportion of the total phenotypic variation attributable to the genetic component, as defined in Rodríguez-Álvarez et al. (2018); this gives geneticists an idea of the signal to noise ratio.

4.1.3 Spatially corrected phenotypic trait

Once the SpATS model in equation (4.2) is fitted, the phenotype of interest at time t, $\boldsymbol{y}_t$, is corrected by only considering the (estimated) sources of variation that are of interest, plus the residual component. In particular, we will retain the (random) genotypic, $\boldsymbol{Z}_{\text{gen}}\boldsymbol{u}_{\text{gen},t}$, and (fixed) population effects, $\boldsymbol{X}_{\text{pop}}\boldsymbol{\beta}_{\text{pop},t}$. Thus, the spatially corrected phenotype at time t, denoted as $\tilde{\boldsymbol{y}}_t = (\tilde{y}_1(t), \ldots, \tilde{y}_M(t))^T$, is obtained as follows

$$\tilde{\boldsymbol{y}}_t = \underbrace{\mathbf{1}_M\hat{\beta}_{0t} + \sum_{e_1=1}^{E_1} \frac{1}{\hat{c}_{e_1}} \boldsymbol{J}_{e_1}\hat{\boldsymbol{\beta}}_{e_1 t} + \boldsymbol{X}_{\text{pop}}\hat{\boldsymbol{\beta}}_{\text{pop},t} + \boldsymbol{Z}_{\text{gen}}\hat{\boldsymbol{u}}_{\text{gen},t}}_{\hat{\boldsymbol{p}}_t} + \hat{\boldsymbol{\varepsilon}}_t, \tag{4.4}$$

where $\boldsymbol{J}_{e_1}$ are matrices of ones of appropriate dimensions (i.e., $M \times (\hat{c}_{e_1} - 1)$), and $\hat{\boldsymbol{p}}_t$ represents predicted values for the genetic populations and the genotypes at time t.

The correction is performed following the procedure for obtaining predictions (e.g., adjusted means) in linear mixed models described in Welham et al. (2004). In that paper, the authors propose a partition of the explanatory variables, e.g., in model (4.2), in three groups:

1. those for which predictions are required (i.e., sources of variation that are of interest for the second stage): population $\boldsymbol{X}_{\text{pop}}\hat{\boldsymbol{\beta}}_{\text{pop},t}$, and genotypic effects $\boldsymbol{Z}_{\text{gen}}\hat{\boldsymbol{u}}_{\text{gen},t}$,

2. those to be ignored: spatial trends, $\hat{\boldsymbol{h}}_{\mathrm{S},t}$, random row and column effects, $\boldsymbol{Z}_{\mathrm{row}}\hat{\boldsymbol{u}}_{\mathrm{row},t}$ and $\boldsymbol{Z}_{\mathrm{col}}\hat{\boldsymbol{u}}_{\mathrm{col},t}$, and other experimental design factors included as random effects in $\boldsymbol{Z}_{\hat{\mathrm{S}}}\hat{\boldsymbol{u}}_{\hat{\mathrm{S}},t}$, and

3. those to be averaged over (i.e., sources of variation that are not of interest for the second stage): $\sum_{e_1=1}^{E_1}(1/\hat{c}_{e_1})\boldsymbol{J}_{e_1}\hat{\boldsymbol{\beta}}_{e_1 t}$, that is experimental design factors included as fixed effect coefficients in $\boldsymbol{X}_{\hat{\mathrm{S}}}\boldsymbol{\beta}_{\hat{\mathrm{S}},t}$ (e.g., presence of block and/or replication and/or lot effects).

As result of the first stage of our two-stage approach we obtain spatially corrected time-series at the resolution of plants or plots with reduced between replicates/plots variability.

4.1.4 Error propagation

For the second stage of our proposal, we model the spatially corrected phenotype of interest $\tilde{\boldsymbol{y}}_t$ ($t \in \{t_1, \ldots, t_n\}$). Thus, it is worth emphasising that, in the way it is constructed, $\tilde{\boldsymbol{y}}_t$ (see equation (4.4)) only contains information about genetic populations and genotypes, as well as unexplained plant-to-plant variation (measurement error). In other words, for the second stage the predicted values for the genetic populations and the genotypes, $\hat{\boldsymbol{p}}_t$, as well as the unexplained plant-to-plant variation, $\hat{\boldsymbol{\varepsilon}}_t$, are maintained as the "new" (spatially corrected) experimental unit values, while the spatial trends and other blocking factors to control for spatial variability are omitted. Also, note that the "observations" that enter the second stage, $\tilde{\boldsymbol{y}}_t$, are not observed but estimated/predicted. Thus, we propose to propagate the uncertainty from the first stage to the second stage through the inclusion of weights, in a similar way to the weighted stage-wise analysis of multi-environment trials (see, e.g., Buntaran et al., 2020). In particular, weights are obtained from the inverse of the variance-covariance (vcov) matrix for the predictions (plus the residual variance), i.e.,

$$\boldsymbol{w}_t = \mathrm{diag}((\mathrm{vcov}(\hat{\boldsymbol{p}}_t) + \hat{\sigma}_t^2 \boldsymbol{I}_M)^{-1}), \tag{4.5}$$

where $\boldsymbol{w}_t = (w_1(t), \ldots, w_M(t))^T$. We note that if the error is not propagated from the first to the second stage, then $\boldsymbol{w}_t = \boldsymbol{1}_M$.

4.1.5 SpATS model estimation and computational aspects

SpATS model in equation (4.2) is a standard linear mixed model, and, thus, estimation is performed as usual in the mixed-model framework. Thus, BLUEs and BLUPs are obtained by the solution of Henderson's mixed model equations (Henderson, 1963), and variance components by means of REML. We use the so-called R-package `SpATS` to estimate model (4.2), which is freely available on CRAN (https://CRAN.R-project.org/package=SpATS), and the recent `statgenHTP` R-package (Millet et al., 2022, available on https://CRAN.R-project.org/package=statgenHTP) that allows for an easy fitting of SpATS models for different (and possibly a large number of) measurement times (through the R-function `fitModels`, and

option `engine = "SpATS"`). For more details in the SpATS model estimation, we refer the reader to Rodríguez-Álvarez et al. (2018). Instead, we will focus on the estimation and computational details of the second stage model (see Section 4.2.5), which is the main contribution of our two-stage approach.

4.2 Second-stage: Temporal evolution of the genetic signal

The aim of the second stage is to model the spatially corrected phenotype obtained in the first stage. We can re-organise the data for this stage in such a way that they can be seen as a sample of plant curves, $\tilde{\boldsymbol{y}}_i = (\tilde{y}_i(t_1), \ldots, \tilde{y}_i(t_n))$ $(1 \leq i \leq M)$, with a nested hierarchical structure, where plants are nested in genotypes and genotypes are nested in populations. We propose to model this sample of curves by considering an additive decomposition of the phenotypic variation over time and use a three-level nested hierarchical model for this purpose (Brumback & Rice, 1998)

$$\tilde{y}_i(t) = f_{p(i)}(t) + f_{g(i)}(t) + f_i(t) + \varepsilon_i(t), \quad \varepsilon_i(t) \sim N\left(0, \sigma^2 w_i(t)\right), \quad 1 \leq p \leq K, \quad 1 \leq g \leq L, \quad 1 \leq i \leq M, \quad (4.6)$$

where f_p is the growth/change over time of the (spatially corrected) phenotype for the pth population (i.e., the pth population mean function), f_g is the genotype-specific deviation from f_p for the gth genotype, and f_i is the plant-specific deviation from f_g for the ith plant. The additive modelling approach implies that $f_p + f_g$ can be interpreted as the evolution over time of the (spatially corrected) phenotype for the gth genotype in the pth population. Thus, on top of genotype-specific deviations from their overall population mean, we also obtain genotype-specific curves. Finally, $w_i(t)$ are the weights as obtained in (4.5) in the first stage of the approach.

4.2.1 P-spline-based hierarchical data model (psHDM)

We use P-splines for hierarchical curve data (Durban et al., 2005; Greven & Scheipl, 2017) to estimate the model in equation (4.6). Henceforth, we call model (4.6) the P-spline-based hierarchical data model (psHDM). In this framework, each function in equation (4.6) is approximated by a linear combination of cubic B-spline basis functions as follows

$$\tilde{y}_i(t) = \underbrace{\sum_{k_{\text{pop}}=1}^{b_{\text{pop}}} B_{k_{\text{pop}}}(t)\theta^{\text{pop}}_{p(i),k_{\text{pop}}}}_{f_{p(i)}(t)} + \underbrace{\sum_{k_{\text{gen}}=1}^{b_{\text{gen}}} B_{k_{\text{gen}}}(t)\theta^{\text{gen}}_{g(i),k_{\text{gen}}}}_{f_{g(i)}(t)} + \underbrace{\sum_{k_{\text{plant}}=1}^{b_{\text{plant}}} B_{k_{\text{plant}}}(t)\theta^{\text{plant}}_{i,k_{\text{plant}}}}_{f_i(t)} + \varepsilon_i(t), \quad (4.7)$$

where $\boldsymbol{\theta}^{\text{pop}}_{p(i)} = (\theta^{\text{pop}}_{p(i),1}, \ldots, \theta^{\text{pop}}_{p(i),b_{\text{pop}}})^T$, $\boldsymbol{\theta}^{\text{gen}}_{g(i)} = (\theta^{\text{gen}}_{g(i),1}, \ldots, \theta^{\text{gen}}_{g(i),b_{\text{gen}}})^T$, and $\boldsymbol{\theta}^{\text{plant}}_{i} = (\theta^{\text{plant}}_{i,1}, \ldots, \theta^{\text{plant}}_{i,b_{\text{plant}}})^T$ are vectors of unknown regression coefficients that control the shape of the curves at the three levels of the hierarchy. If $\tilde{\boldsymbol{y}}_i = (\tilde{y}_i(t_1), \ldots, \tilde{y}_i(t_n))^T$ represents the measurements for a single plant, in matrix form model (4.7) becomes

$$\tilde{\boldsymbol{y}}_i = \underbrace{\boldsymbol{B}_{\text{pop}}\boldsymbol{\theta}^{\text{pop}}_{p(i)}}_{\boldsymbol{f}_p} + \underbrace{\boldsymbol{B}_{\text{gen}}\boldsymbol{\theta}^{\text{gen}}_{g(i)}}_{\boldsymbol{f}_g} + \underbrace{\boldsymbol{B}_{\text{plant}}\boldsymbol{\theta}^{\text{plant}}_{i}}_{\boldsymbol{f}_i} + \boldsymbol{\varepsilon}_i, \tag{4.8}$$

where $\boldsymbol{f}_p = (f_{p(i)}(t_1), \ldots, f_{p(i)}(t_n))^T$, $\boldsymbol{f}_g = (f_{g(i)}(t_1), \ldots, f_{g(i)}(t_n))^T$, $\boldsymbol{f}_i = (f_i(t_1), \ldots, f_i(t_n))^T$, and $(\boldsymbol{B}_{\text{pop}})^{n\times b_{\text{pop}}}_{jk_{\text{pop}}} = B_{k_{\text{pop}}}(t_j)$, $(\boldsymbol{B}_{\text{geno}})^{n\times b_{\text{gen}}}_{jk_{\text{gen}}} = B_{k_{\text{gen}}}(t_j)$ and $(\boldsymbol{B}_{\text{plant}})^{n\times b_{\text{plant}}}_{jk_{\text{plant}}} = B_{k_{\text{plant}}}(t_j)$. As usual in P-splines, smoothness is controlled by a penalty on the differences of the B-splines coefficients (we use differences of order 2). The influence of the penalty is determined by the smoothing parameters. For the hierarchical model (4.8), we have an additive penalty of the form

$$\lambda_{\text{pop},p}\boldsymbol{\theta}^{\text{pop};T}_{p(i)}\boldsymbol{P}_{\text{pop}}\boldsymbol{\theta}^{\text{pop}}_{p(i)} + \lambda_{\text{gen}}\boldsymbol{\theta}^{\text{gen};T}_{g(i)}\boldsymbol{P}_{\text{gen}}\boldsymbol{\theta}^{\text{gen}}_{g(i)} + \lambda_{\text{plant}}\boldsymbol{\theta}^{\text{plant};T}_{i}\boldsymbol{P}_{\text{plant}}\boldsymbol{\theta}^{\text{plant}}_{i}, \tag{4.9}$$

where $\boldsymbol{P}_\nu = \boldsymbol{D}^T_\nu\boldsymbol{D}_\nu$ ($\nu \in \{\text{pop}, \text{gen}, \text{plant}\}$) are penalty matrices with $\boldsymbol{D}_\nu$ matrices that form second order differences at each level of the hierarchy, and $\lambda_{\text{pop},p}$, λ_{gen} and λ_{plant} are the smoothing parameters.

We finally present model (4.8) for all plants. Let's first order the data by plant, and time, i.e., $\tilde{\boldsymbol{y}} = (\tilde{y}_1(t_1), \ldots, \tilde{y}_1(t_n), \ldots, \tilde{y}_M(t_1), \ldots, \tilde{y}_M(t_n))^T$. In fact, in order for the following Kronecker products to make sense, the data should be pre-ordered by population, genotype, plant and time (in that order), where the first m_1 plants belongs to the first genotype (i.e., $g = 1$) of the first population (i.e., $p = 1$). Thus, in a compact way, the three-level nested hierarchical model can be expressed as

$$\tilde{\boldsymbol{y}} = (\boldsymbol{Q}_{\text{pop}} \otimes \boldsymbol{B}_{\text{pop}})\boldsymbol{\theta}_{\text{pop}} + (\boldsymbol{Q}_{\text{gen}} \otimes \boldsymbol{B}_{\text{gen}})\boldsymbol{\theta}_{\text{gen}} + (\boldsymbol{I}_M \otimes \boldsymbol{B}_{\text{plant}})\boldsymbol{\theta}_{\text{plant}} + \boldsymbol{\varepsilon}, \tag{4.10}$$

where $\otimes$ is the Kronecker product, and $\boldsymbol{Q}_{\text{pop}}$ and $\boldsymbol{Q}_{\text{gen}}$ are matrices assigning, respectively, plants to populations and plants to genotypes. That is, $\boldsymbol{Q}^{M\times K}_{\text{pop}} = \text{blockdiag}(\boldsymbol{1}^{\text{pop};T}_1, \ldots, \boldsymbol{1}^{\text{pop};T}_K)$, with $\boldsymbol{1}^{\text{pop}}_p$ vectors of ones with appropriate length (i.e., $\#\{i \mid p(i) = p\}$), and $\boldsymbol{Q}^{M\times L}_{\text{gen}} = \text{blockdiag}(\boldsymbol{1}^{\text{gen};T}_1, \ldots, \boldsymbol{1}^{\text{gen};T}_L)$, with $\boldsymbol{1}^{\text{gen}}_g$ vectors of ones of length m_g, and $\boldsymbol{\theta}_{\text{pop}} = (\boldsymbol{\theta}^{\text{pop};T}_1, \ldots, \boldsymbol{\theta}^{\text{pop};T}_K)^T$, $\boldsymbol{\theta}_{\text{gen}} = (\boldsymbol{\theta}^{\text{gen};T}_1, \ldots, \boldsymbol{\theta}^{\text{gen};T}_L)^T$, and $\boldsymbol{\theta}_{\text{plant}} = (\boldsymbol{\theta}^{\text{plant};T}_1, \ldots, \boldsymbol{\theta}^{\text{plant};T}_M)^T$ vectors of unknown regression coefficients. The penalty associated with model (4.7) is

$$\sum_{p=1}^{K} \lambda_{\text{pop},p}\boldsymbol{\theta}^{\text{pop};T}_p\boldsymbol{P}_{\text{pop}}\boldsymbol{\theta}^{\text{pop}}_p + \lambda_{\text{gen}}\boldsymbol{\theta}^T_{\text{gen}}(\boldsymbol{I}_L \otimes \boldsymbol{P}_{\text{gen}})\boldsymbol{\theta}_{\text{gen}} + \lambda_{\text{plant}}\boldsymbol{\theta}^T_{\text{plant}}(\boldsymbol{I}_M \otimes \boldsymbol{P}_{\text{plant}})\boldsymbol{\theta}_{\text{plant}}, \tag{4.11}$$

That is, we allow for different smoothing parameters for curves at the population level (i.e., $\boldsymbol{\lambda}_{\text{pop}} = (\lambda_{\text{pop},1}, \ldots, \lambda_{\text{pop},K})$), while keeping constant the smoothing parameter for all genotypic and plant curves.

4.2.2 Mixed model formulation of the psHDM

To find appropriate values for the smoothing parameters $\boldsymbol{\lambda}_{\text{pop}}$, λ_{gen}, and λ_{plant}, we adopt the connection between P-splines and linear mixed models in a similar fashion to the procedures described in Section 3.2, where each smooth function is treated as a sum of fixed (linear) and random (non-linear) components, and

the smoothing parameters are replaced by a ratio of variances components. Under this framework, the mixed model representation of the P-spline model (4.8) for one plant is

$$\tilde{\boldsymbol{y}}_i = \underbrace{\boldsymbol{X}_{\text{pop}}\boldsymbol{\beta}_{\text{lin},p(i)} + \boldsymbol{Z}_{\text{pop}}\boldsymbol{u}^{\text{pop}}_{\text{nlin},p(i)}}_{f_p} + \underbrace{\boldsymbol{X}_{\text{gen}}\boldsymbol{u}^{\text{gen}}_{\text{lin},g(i)} + \boldsymbol{Z}_{\text{gen}}\boldsymbol{u}^{\text{gen}}_{\text{nlin},g(i)}}_{f_g} + \underbrace{\boldsymbol{X}_{\text{plant}}\boldsymbol{u}^{\text{plant}}_{\text{lin},i} + \boldsymbol{Z}_{\text{plant}}\boldsymbol{u}^{\text{plant}}_{\text{nlin},i}}_{f_i} + \boldsymbol{\varepsilon}_i,\ \boldsymbol{\varepsilon}_i \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{w}_i), \tag{4.12}$$

To obtain expressions for the mixed model design matrices, for the regression coefficients, and for the variance-covariance matrix associated with the random effects, we follow the ideas described in Section 3.2.1. That is, we use the SVD of the penalty matrices $\boldsymbol{P}_\nu = \boldsymbol{D}_\nu^T\boldsymbol{D}_\nu = \boldsymbol{U}_\nu\boldsymbol{\Lambda}_\nu\boldsymbol{U}_\nu^T$ ($\nu \in \{\text{pop}, \text{gen}, \text{plant}\}$). As before, $\boldsymbol{U}_\nu$ is the matrix of eigen vectors and $\boldsymbol{\Lambda}_\nu$ is the diagonal matrix of eigenvalues, such that $\boldsymbol{U}_{\nu+}$ ($\boldsymbol{\Lambda}_{\nu+}$) and $\boldsymbol{U}_{\nu 0}$ ($\boldsymbol{\Lambda}_{\nu 0}$) are the submatrices corresponding, respectively, to the non-zero and zero eigenvalues. Therefore, the specification of model (4.12) is as follows

$$\begin{aligned}
\boldsymbol{X}_{\text{pop}} &= \boldsymbol{B}_{\text{pop}}\boldsymbol{U}_{\text{pop},0} \text{ and } \boldsymbol{\beta}^{\text{pop}}_{\text{lin},p} = \boldsymbol{U}^T_{\text{pop},0}\boldsymbol{\theta}^{\text{pop}}_p,\\
\boldsymbol{Z}_{\text{pop}} &= \boldsymbol{B}_{\text{pop}}\boldsymbol{U}_{\text{pop}+} \text{ and } \boldsymbol{u}^{\text{pop}}_{\text{nlin},p} = \boldsymbol{U}^T_{\text{pop}+}\boldsymbol{\theta}^{\text{pop}}_p = (u^{\text{pop}}_{\text{nlin},p,1}, \ldots, u^{\text{pop}}_{\text{nlin},p,b_{\text{pop}}-2})^T \sim N(\boldsymbol{0}, \sigma^2_{\text{pop},p}\boldsymbol{\Sigma}_{\text{pop}+}),\\
\boldsymbol{X}_{\text{gen}} &= \boldsymbol{B}_{\text{gen}}\boldsymbol{U}_{\text{gen},0} \text{ and } \boldsymbol{\beta}^{\text{gen}}_{\text{lin},g} = \boldsymbol{U}^T_{\text{gen},0}\boldsymbol{\theta}^{\text{gen}}_g,\\
\boldsymbol{Z}_{\text{gen}} &= \boldsymbol{B}_{\text{gen}}\boldsymbol{U}_{\text{gen}+} \text{ and } \boldsymbol{u}^{\text{gen}}_{\text{nlin},g} = \boldsymbol{U}^T_{\text{gen}+}\boldsymbol{\theta}^{\text{gen}}_g = (u^{\text{gen}}_{\text{nlin},g,1}, \ldots, u^{\text{gen}}_{\text{nlin},g,b_{\text{gen}}-2})^T \sim N(\boldsymbol{0}, \sigma^2_{\text{gen}}\boldsymbol{\Sigma}_{\text{gen}+}),\\
\boldsymbol{X}_{\text{plant}} &= \boldsymbol{B}_{\text{plant}}\boldsymbol{U}_{\text{plant},0} \text{ and } \boldsymbol{\beta}^{\text{plant}}_{\text{lin},i} = \boldsymbol{U}^T_{\text{plant},0}\boldsymbol{\theta}^{\text{plant}}_i,\\
\boldsymbol{Z}_{\text{plant}} &= \boldsymbol{B}_{\text{plant}}\boldsymbol{U}_{\text{plant}+} \text{ and } \boldsymbol{u}^{\text{plant}}_{\text{nlin},i} = \boldsymbol{U}^T_{\text{plant}+}\boldsymbol{\theta}^{\text{plant}}_i = (u^{\text{plant}}_{\text{nlin},i,1}, \ldots, u^{\text{plant}}_{\text{nlin},i,b_{\text{plant}}-2})^T \sim N(\boldsymbol{0}, \sigma^2_{\text{plant}}\boldsymbol{\Sigma}_{\text{plant}+}),
\end{aligned}$$

where the variance-covariance matrices for $\boldsymbol{u}^{\text{pop}}_{\text{nlin},p}$, $\boldsymbol{u}^{\text{gen}}_{\text{nlin},g}$, and $\boldsymbol{u}^{\text{plant}}_{\text{nlin},i}$ are obtained from the inverse of their respective precision matrices. That is, $\sigma^2_\nu\boldsymbol{\Sigma}_{\nu+} = \sigma^2/\lambda_\nu\boldsymbol{\Lambda}^{-1}_{\nu+}$, with $\sigma^2_\nu = \sigma^2/\lambda_\nu$ and $\boldsymbol{\Sigma}_{\nu+} = \boldsymbol{\Lambda}^{-1}_{\nu+}$ ($\nu \in \{\text{gen}, \text{plant}\}$). A similar reasoning is followed at the population level but with different smoothing/variance parameters for each population. That is, $\sigma^2_{\text{pop},p}\boldsymbol{\Sigma}_{\text{pop}+} = \sigma^2/\lambda_{\text{pop},p}\boldsymbol{\Lambda}^{-1}_{\text{pop}+}$, with $\sigma^2_{\text{pop},p} = \sigma^2/\lambda_{\text{pop},p}$ and $\boldsymbol{\Sigma}_{\text{pop}+} = \boldsymbol{\Lambda}^{-1}_{\text{pop}+}$. Finally, $\boldsymbol{w}_i$ in (4.12) is a diagonal matrix whose diagonal entries are the weights from the first stage for the ith plant.

In addition, note that based on the SVD, $\boldsymbol{X}_\nu = \boldsymbol{B}_\nu\boldsymbol{U}_{\nu 0}$ ($\nu \in \{\text{pop}, \text{gen}, \text{plant}\}$). However, for simplicity we consider the parameterisation $\boldsymbol{X}_\nu = [\boldsymbol{1}_n \mid \boldsymbol{t}]$ with $\boldsymbol{\beta}^\nu_{\text{lin},\hat{\nu}} = (\beta^\nu_{\text{lin},\hat{\nu},0}, \beta^\nu_{\text{lin},\hat{\nu},1})^T$, $\hat{\nu} \in \{p, g, i\}$, the corresponding intercept and slope. Moreover, in contrast to standard P-spline mixed models, but in line with the traditional random intercept and slope model for longitudinal data, here the linear components (intercept and slope) associated with $\boldsymbol{f}_g$ (genotypic deviations) and $\boldsymbol{f}_i$ (plant deviations) are modelled with penalised/random rather than unpenalised/fixed effect coefficients. There are two reasons for this decision. On the one hand, we treat genotypes and plants as random samples. On the other hand, it avoids identifiability problems that arise with fixed effect coefficients for nested ANOVA models. (Brumback & Rice, 1998) Thus, these

components of model (4.12) become

$$\begin{aligned}
\boldsymbol{X}_{\text{pop}} &= [\mathbf{1}_n \mid \boldsymbol{t}] \text{ and } \boldsymbol{\beta}_{\text{lin},p}^{\text{pop}} = (\beta_{\text{lin},p,0}^{\text{pop}}, \beta_{\text{lin},p,1}^{\text{pop}})^T, \\
\boldsymbol{X}_{\text{gen}} &= [\mathbf{1}_n \mid \boldsymbol{t}] \text{ and } \boldsymbol{\beta}_{\text{lin},g}^{\text{gen}} = \boldsymbol{u}_{\text{lin},g}^{\text{gen}} = (u_{\text{lin},g,0}^{\text{gen}}, u_{\text{lin},g,1}^{\text{gen}})^T \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{\text{gen}}) \text{ with } \boldsymbol{\Sigma}_{\text{gen}} = \text{diag}(\sigma_{\text{gen},0}^2, \sigma_{\text{gen},1}^2), \\
\boldsymbol{X}_{\text{plant}} &= [\mathbf{1}_n \mid \boldsymbol{t}] \text{ and } \boldsymbol{\beta}_{\text{lin},i}^{\text{plant}} = \boldsymbol{u}_{\text{lin},i}^{\text{plant}} = (u_{\text{lin},i,0}^{\text{plant}}, u_{\text{lin},i,1}^{\text{plant}})^T \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{\text{plant}}) \text{ with } \boldsymbol{\Sigma}_{\text{plant}} = \text{diag}(\sigma_{\text{plant},0}^2, \sigma_{\text{plant},1}^2).
\end{aligned}$$

We now present the mixed model (4.12) for all plants, as we did before for the P-spline model (4.10)

$$\tilde{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\varepsilon},\ \boldsymbol{u} \sim N(\mathbf{0}, \boldsymbol{G}),\ \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{W}), \tag{4.13}$$

where,

$$\begin{aligned}
\boldsymbol{X} &= [\boldsymbol{Q}_{\text{pop}} \otimes \boldsymbol{X}_{\text{pop}}], \\
\boldsymbol{Z} &= [\boldsymbol{Q}_{\text{pop}} \otimes \boldsymbol{Z}_{\text{pop}} \mid \boldsymbol{Q}_{\text{gen}} \otimes \boldsymbol{X}_{\text{gen}} \mid \boldsymbol{Q}_{\text{gen}} \otimes \boldsymbol{Z}_{\text{gen}} \mid \boldsymbol{I}_M \otimes \boldsymbol{X}_{\text{plant}} \mid \boldsymbol{I}_M \otimes \boldsymbol{Z}_{\text{plant}}],
\end{aligned}$$

with, $\boldsymbol{\beta}_{\text{pop}} = (\boldsymbol{\beta}_{\text{lin},1}^{\text{pop};T}, \ldots, \boldsymbol{\beta}_{\text{lin},K}^{\text{pop};T})^T$ and $\boldsymbol{u} = (\boldsymbol{u}_{\text{nlin,pop}}^T, \boldsymbol{u}_{\text{lin,gen}}^T, \boldsymbol{u}_{\text{nlin,gen}}^T, \boldsymbol{u}_{\text{lin,plant}}^T, \boldsymbol{u}_{\text{nlin,plant}}^T)^T$, where

$$\begin{aligned}
\boldsymbol{u}_{\text{nlin,pop}} &= (\boldsymbol{u}_{\text{nlin},1}^{\text{pop};T}, \ldots, \boldsymbol{u}_{\text{nlin},K}^{\text{pop};T})^T \sim N(\mathbf{0}, \text{blockdiag}(\sigma_{\text{pop},1}^2 \boldsymbol{\Sigma}_{\text{pop+}}, \ldots, \sigma_{\text{pop},K}^2 \boldsymbol{\Sigma}_{\text{pop+}}), \\
\boldsymbol{u}_{\text{lin,gen}} &= (\boldsymbol{u}_{\text{lin},1}^{\text{gen};T}, \ldots, \boldsymbol{u}_{\text{lin},L}^{\text{gen};T})^T \sim N(\mathbf{0}, \boldsymbol{I}_L \otimes \boldsymbol{\Sigma}_{\text{gen}}), \\
\boldsymbol{u}_{\text{nlin,gen}} &= (\boldsymbol{u}_{\text{nlin},1}^{\text{gen};T}, \ldots, \boldsymbol{u}_{\text{nlin},L}^{\text{gen};T})^T \sim N(\mathbf{0}, \sigma_{\text{gen}}^2 \boldsymbol{I}_L \otimes \boldsymbol{\Sigma}_{\text{gen+}}), \\
\boldsymbol{u}_{\text{lin,plant}} &= (\boldsymbol{u}_{\text{lin},1}^{\text{plant};T}, \ldots, \boldsymbol{u}_{\text{lin},M}^{\text{plant};T})^T \sim N(\mathbf{0}, \boldsymbol{I}_M \otimes \boldsymbol{\Sigma}_{\text{plant}}), \\
\boldsymbol{u}_{\text{nlin,plant}} &= (\boldsymbol{u}_{\text{nlin},1}^{\text{plant};T}, \ldots, \boldsymbol{u}_{\text{nlin},M}^{\text{plant};T})^T \sim N(\mathbf{0}, \sigma_{\text{plant}}^2 \boldsymbol{I}_M \otimes \boldsymbol{\Sigma}_{\text{plant+}}),
\end{aligned}$$

and the variance-covariance matrix for random effects $\boldsymbol{u}$

$$\begin{aligned}
\boldsymbol{G} &= \begin{pmatrix}
\text{blockdiag}(\sigma_{\text{pop},1}^2 \boldsymbol{\Sigma}_{\text{pop+}}, \ldots, \sigma_{\text{pop},K}^2 \boldsymbol{\Sigma}_{\text{pop+}}) & \mathbf{0} & \mathbf{0} \\
\mathbf{0} & \begin{pmatrix} \boldsymbol{I}_L \otimes \boldsymbol{\Sigma}_{\text{gen}} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I}_L \otimes \sigma_{\text{gen}}^2 \boldsymbol{\Sigma}_{\text{gen+}} \end{pmatrix} & \mathbf{0} \\
\mathbf{0} & \mathbf{0} & \begin{pmatrix} \boldsymbol{I}_M \otimes \boldsymbol{\Sigma}_{\text{plant}} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I}_M \otimes \sigma_{\text{plant}}^2 \boldsymbol{\Sigma}_{\text{plant+}} \end{pmatrix}
\end{pmatrix} \\
&= \sum_{j=1}^{K+6} \sigma_j^2 \tilde{\boldsymbol{\Sigma}}_j,
\end{aligned} \tag{4.14}$$

where

$$
\begin{aligned}
\tilde{\boldsymbol{\Sigma}}_1 &= \text{blockdiag}(\boldsymbol{\Sigma}_{\text{pop+}}, \ldots, \mathbf{0}, \boldsymbol{I}_L \otimes \text{blockdiag}(0,0,\mathbf{0}), \boldsymbol{I}_M \otimes \text{blockdiag}(0,0,\mathbf{0})),\ \sigma^2_1 = \sigma^2_{\text{pop},1} \\
&\vdots \\
\tilde{\boldsymbol{\Sigma}}_K &= \text{blockdiag}(\mathbf{0}, \ldots, \boldsymbol{\Sigma}_{\text{pop+}}, \boldsymbol{I}_L \otimes \text{blockdiag}(0,0,\mathbf{0}), \boldsymbol{I}_M \otimes \text{blockdiag}(0,0,\mathbf{0})),\ \sigma^2_K = \sigma^2_{\text{pop},K} \\
\tilde{\boldsymbol{\Sigma}}_{K+1} &= \text{blockdiag}(\mathbf{0}, \ldots, \mathbf{0}, \boldsymbol{I}_L \otimes \text{blockdiag}(1,0,\mathbf{0}), \boldsymbol{I}_M \otimes \text{blockdiag}(0,0,\mathbf{0})),\ \sigma^2_{K+1} = \sigma^2_{\text{gen},0}, \\
\tilde{\boldsymbol{\Sigma}}_{K+2} &= \text{blockdiag}(\mathbf{0}, \ldots, \mathbf{0}, \boldsymbol{I}_L \otimes \text{blockdiag}(0,1,\mathbf{0}), \boldsymbol{I}_M \otimes \text{blockdiag}(0,0,\mathbf{0})),\ \sigma^2_{K+2} = \sigma^2_{\text{gen},1}, \\
\tilde{\boldsymbol{\Sigma}}_{K+3} &= \text{blockdiag}(\mathbf{0}, \ldots, \mathbf{0}, \boldsymbol{I}_L \otimes \text{blockdiag}(0,0,\boldsymbol{\Sigma}_{\text{gen+}}), \boldsymbol{I}_M \otimes \text{blockdiag}(0,0,\mathbf{0})),\ \sigma^2_{K+3} = \sigma^2_{\text{gen}}, \\
\tilde{\boldsymbol{\Sigma}}_{K+4} &= \text{blockdiag}(\mathbf{0}, \ldots, \mathbf{0}, \boldsymbol{I}_L \otimes \text{blockdiag}(0,0,\mathbf{0}), \boldsymbol{I}_M \otimes \text{blockdiag}(1,0,\mathbf{0})),\ \sigma^2_{K+4} = \sigma^2_{\text{plant},0}, \\
\tilde{\boldsymbol{\Sigma}}_{K+5} &= \text{blockdiag}(\mathbf{0}, \ldots, \mathbf{0}, \boldsymbol{I}_L \otimes \text{blockdiag}(0,0,\mathbf{0}), \boldsymbol{I}_M \otimes \text{blockdiag}(0,1,\mathbf{0})),\ \sigma^2_{K+5} = \sigma^2_{\text{plant},1}, \\
\tilde{\boldsymbol{\Sigma}}_{K+6} &= \text{blockdiag}(\mathbf{0}, \ldots, \mathbf{0}, \boldsymbol{I}_L \otimes \text{blockdiag}(0,0,\mathbf{0}), \boldsymbol{I}_M \otimes \text{blockdiag}(0,0,\boldsymbol{\Sigma}_{\text{plant+}})),\ \sigma^2_{K+6} = \sigma^2_{\text{plant}}.
\end{aligned}
$$

Finally, we comment on the selection of the number of B-spline basis functions used to approximate f_p, f_g and f_i (i.e., b_{pop}, b_{gen} and b_{plant}, respectively). In P-splines, it is recommended to choose a large number of bases to provide enough flexibility; the role of the penalty is to avoid over fitting (Eilers & Marx, 1996). In our setting the number of functions in the complete model equals $K + L + M$ (populations + genotypes + plants), and, hence, the number of regression coefficients (either fixed or random) to be estimated is $K \times b_{\text{pop}} + L \times b_{\text{gen}} + M \times b_{\text{plant}}$. This value can be very large, with the number of plants, M, and associated basis dimension, b_{plant}, playing the major role: the dataset may contain thousands of plants. Thus, to reduce the computational burden, one could be tempted to use different basis dimensions for f_p, f_g and f_i, and to be less generous with f_i. However, this is not a good strategy. As will be shown later, simulation studies, as well as preparatory data analyses, have shown that results may be sensitive (and in some cases unreliable) to using different bases dimensions. We therefore recommend choosing the same value for b_{pop}, b_{gen} and b_{plant}, while keeping the number of coefficients at a reasonable level (i.e., a trade-off between flexibility and dimensionality). In addition to the number of regression coefficients, under model (4.13), the number of variance components to estimate is: K at population level ($\sigma^2_{\text{pop},1}, \ldots, \sigma^2_{\text{pop},K}$), 3 at genotype level ($\sigma^2_{\text{gen},0}, \sigma^2_{\text{gen},1}, \sigma^2_{\text{gen}}$), 3 at plant level ($\sigma^2_{\text{plant},0}, \sigma^2_{\text{plant},1}, \sigma^2_{\text{plant}}$), and the error variance (σ^2).

4.2.3 psHDM with different genetic and/or plant-to-plant variation

In model (4.13) we assume the same genetic variation across populations (see variance-covariance matrix specification, $\boldsymbol{G}$ in (4.14)). However, this assumption can be easily relaxed by considering different values of $\sigma^2_{\text{gen},0}$, $\sigma^2_{\text{gen},1}$ and σ^2_{gen} per population. A similar approach can be followed to allow for the plant-to-plant variation ($\sigma^2_{\text{plant},0}$, $\sigma^2_{\text{plant},1}$ and σ^2_{plant}) to vary across genotypes. These generalisations might be worth exploring if there are sufficient number of genotypes per population and plants per genotype, respectively.

For this purpose, the following specification of the model (4.13) has to be made

$$
\begin{aligned}
\boldsymbol{u}_{\text{nlin,pop}} &= (\boldsymbol{u}_{\text{nlin},1}^{\text{pop};T}, \ldots, \boldsymbol{u}_{\text{nlin},K}^{\text{pop};T})^T \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_{\text{pop}}) \text{ with } \boldsymbol{\Sigma}_{\text{pop}} = \text{blockdiag}(\sigma^2_{\text{pop},1}\boldsymbol{\Sigma}_{\text{pop+}}, \ldots, \sigma^2_{\text{pop},K}\boldsymbol{\Sigma}_{\text{pop+}}),\\
\boldsymbol{u}_{\text{lin,gen}} &= (\boldsymbol{u}_{\text{lin},1}^{\text{gen};T}, \ldots, \boldsymbol{u}_{\text{lin},L}^{\text{gen};T})^T \sim N(\boldsymbol{0}, \text{blockdiag}(\boldsymbol{I}_{\ell_1} \otimes \boldsymbol{\Sigma}_1^{\text{gen}}, \ldots, \boldsymbol{I}_{\ell_K} \otimes \boldsymbol{\Sigma}_K^{\text{gen}})) \text{ with}\\
&\quad \boldsymbol{\Sigma}_p^{\text{gen}} = \text{diag}(\sigma^2_{\text{gen},p,0}, \sigma^2_{\text{gen},p,1}),\\
\boldsymbol{u}_{\text{nlin,gen}} &= (\boldsymbol{u}_{\text{nlin},1}^{\text{gen};T}, \ldots, \boldsymbol{u}_{\text{nlin},L}^{\text{gen};T})^T \sim N(\boldsymbol{0}, \text{blockdiag}(\sigma^2_{\text{gen},1}\boldsymbol{I}_{\ell_1} \otimes \boldsymbol{\Sigma}_{\text{gen+}}, \ldots, \sigma^2_{\text{gen},K}\boldsymbol{I}_{\ell_K} \otimes \boldsymbol{\Sigma}_{\text{gen+}})),\\
\boldsymbol{u}_{\text{lin,plant}} &= (\boldsymbol{u}_{\text{lin},1}^{\text{plant};T}, \ldots, \boldsymbol{u}_{\text{lin},M}^{\text{plant};T})^T \sim N(\boldsymbol{0}, \text{blockdiag}(\boldsymbol{I}_{m_1} \otimes \boldsymbol{\Sigma}_1^{\text{plant}}, \ldots, \boldsymbol{I}_{m_L} \otimes \boldsymbol{\Sigma}_L^{\text{plant}})) \text{ with}\\
&\quad \boldsymbol{\Sigma}_g^{\text{plant}} = \text{diag}(\sigma^2_{\text{plant},g,0}, \sigma^2_{\text{plant},g,1}),\\
\boldsymbol{u}_{\text{nlin,plant}} &= (\boldsymbol{u}_{\text{nlin},1}^{\text{plant};T}, \ldots, \boldsymbol{u}_{\text{nlin},M}^{\text{plant};T})^T \sim N(\boldsymbol{0}, \text{blockdiag}(\sigma^2_{\text{plant},1}\boldsymbol{I}_{m_1} \otimes \boldsymbol{\Sigma}_{\text{plant+}}, \ldots, \sigma^2_{\text{plant},L}\boldsymbol{I}_{m_L} \otimes \boldsymbol{\Sigma}_{\text{plant+}})),
\end{aligned}
$$

where the variance-covariance matrix for the random effects $\boldsymbol{u}$ is

$$
\boldsymbol{G} = \text{blockdiag}(\boldsymbol{\Sigma}_{\text{pop}}, \boldsymbol{\Sigma}_{\text{gen}}, \boldsymbol{\Sigma}_{\text{plant}}), \tag{4.15}
$$

where

$$
\boldsymbol{\Sigma}_{\text{gen}} = \begin{pmatrix}
\begin{pmatrix} \boldsymbol{I}_{\ell_1} \otimes \boldsymbol{\Sigma}_1^{\text{gen}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{\ell_1} \otimes \sigma^2_{\text{gen},1}\boldsymbol{\Sigma}_{\text{gen+}} \end{pmatrix} & \cdots & \boldsymbol{0} \\
\vdots & \ddots & \vdots \\
\boldsymbol{0} & \cdots & \begin{pmatrix} \boldsymbol{I}_{\ell_K} \otimes \boldsymbol{\Sigma}_K^{\text{gen}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{\ell_K} \otimes \sigma^2_{\text{gen},K}\boldsymbol{\Sigma}_{\text{gen+}} \end{pmatrix}
\end{pmatrix},
$$

and

$$
\boldsymbol{\Sigma}_{\text{plant}} = \begin{pmatrix}
\begin{pmatrix} \boldsymbol{I}_{m_1} \otimes \boldsymbol{\Sigma}_1^{\text{plant}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{m_1} \otimes \sigma^2_{\text{plant},1}\boldsymbol{\Sigma}_{\text{plant+}} \end{pmatrix} & \cdots & \boldsymbol{0} \\
\vdots & \ddots & \vdots \\
\boldsymbol{0} & \cdots & \begin{pmatrix} \boldsymbol{I}_{m_L} \otimes \boldsymbol{\Sigma}_L^{\text{plant}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{m_L} \otimes \sigma^2_{\text{plant},L}\boldsymbol{\Sigma}_{\text{plant+}} \end{pmatrix}
\end{pmatrix},
$$

thus, $\boldsymbol{G}$ can be written as $\boldsymbol{G} = \sum_{j=1}^{K+3K+3L} \sigma_j^2 \tilde{\boldsymbol{\Sigma}}_j$, where

$$
\begin{aligned}
\tilde{\boldsymbol{\Sigma}}_1 = \ & \text{blockdiag}(\boldsymbol{\Sigma}_{\text{pop+}}, \ldots, \mathbf{0}, \\
& \boldsymbol{I}_{\ell_1} \otimes \text{blockdiag}(0, 0, \mathbf{0}), \ldots, \boldsymbol{I}_{\ell_K} \otimes \text{blockdiag}(0, 0, \mathbf{0}), \\
& \boldsymbol{I}_{m_1} \otimes \text{blockdiag}(0, 0, \mathbf{0}), \ldots, \boldsymbol{I}_{m_L} \otimes \text{blockdiag}(0, 0, \mathbf{0})), \\
& \vdots \\
\tilde{\boldsymbol{\Sigma}}_K = \ & \text{blockdiag}(\mathbf{0}, \ldots, \boldsymbol{\Sigma}_{\text{pop+}}, \\
& \boldsymbol{I}_{\ell_1} \otimes \text{blockdiag}(0, 0, \mathbf{0}), \ldots, \boldsymbol{I}_{\ell_K} \otimes \text{blockdiag}(0, 0, \mathbf{0}), \\
& \boldsymbol{I}_{m_1} \otimes \text{blockdiag}(0, 0, \mathbf{0}), \ldots, \boldsymbol{I}_{m_L} \otimes \text{blockdiag}(0, 0, \mathbf{0})), \\
\tilde{\boldsymbol{\Sigma}}_{K+1} = \ & \text{blockdiag}(\mathbf{0}, \ldots, \mathbf{0}, \\
& \boldsymbol{I}_{\ell_1} \otimes \text{blockdiag}(1, 0, \mathbf{0}), \ldots, \boldsymbol{I}_{\ell_K} \otimes \text{blockdiag}(0, 0, \mathbf{0}), \\
& \boldsymbol{I}_{m_1} \otimes \text{blockdiag}(0, 0, \mathbf{0}), \ldots, \boldsymbol{I}_{m_L} \otimes \text{blockdiag}(0, 0, \mathbf{0})), \\
& \vdots \\
\tilde{\boldsymbol{\Sigma}}_{K+3K} = \ & \text{blockdiag}(\mathbf{0}, \ldots, \mathbf{0}, \\
& \boldsymbol{I}_{\ell_1} \otimes \text{blockdiag}(0, 0, \mathbf{0}), \ldots \boldsymbol{I}_{\ell_K} \otimes \text{blockdiag}(0, 0, \boldsymbol{\Sigma}_{\text{gen+}}), \\
& \boldsymbol{I}_{m_1} \otimes \text{blockdiag}(0, 0, \mathbf{0}), \ldots, \boldsymbol{I}_{m_L} \otimes \text{blockdiag}(0, 0, \mathbf{0})), \\
\\
\tilde{\boldsymbol{\Sigma}}_{K+3K+1} = \ & \text{blockdiag}(\mathbf{0}, \ldots, \mathbf{0}, \\
& \boldsymbol{I}_{\ell_1} \otimes \text{blockdiag}(0, 0, \mathbf{0}), \ldots, \boldsymbol{I}_{\ell_K} \otimes \text{blockdiag}(0, 0, \mathbf{0}), \\
& \boldsymbol{I}_{m_1} \otimes \text{blockdiag}(1, 0, \mathbf{0}), \ldots, \boldsymbol{I}_{m_L} \otimes \text{blockdiag}(0, 0, \mathbf{0})), \\
& \vdots \\
\tilde{\boldsymbol{\Sigma}}_{K+3K+3L} = \ & \text{blockdiag}(\mathbf{0}, \ldots, \mathbf{0}, \\
& \boldsymbol{I}_{\ell_1} \otimes \text{blockdiag}(0, 0, \mathbf{0}), \ldots, \boldsymbol{I}_{\ell_K} \otimes \text{blockdiag}(0, 0, \mathbf{0}), \\
& \boldsymbol{I}_{m_1} \otimes \text{blockdiag}(0, 0, \mathbf{0}), \ldots, \boldsymbol{I}_{m_L} \otimes \text{blockdiag}(0, 0, \boldsymbol{\Sigma}_{\text{plant+}})),
\end{aligned}
$$

and variance parameters at genotype level $\sigma^2_{K+1} = \sigma^2_{\text{gen},1,0}$, $\sigma^2_{K+2} = \sigma^2_{\text{gen},1,1}$, $\sigma^2_{K+3} = \sigma^2_{\text{gen},1}$, …, $\sigma^2_{K+3K} = \sigma^2_{\text{gen},K}$, and at plant level $\sigma^2_{K+3K+1} = \sigma^2_{\text{plant},1,0}$, $\sigma^2_{K+3K+2} = \sigma^2_{\text{plant},1,1}$, $\sigma^2_{K+3K+3} = \sigma^2_{\text{plant},1}$, …, $\sigma^2_{K+3K+3L} = \sigma^2_{\text{plant},L}$. We note that, under this configuration of model (4.13), the number of variance components to estimate is: K at population level ($\sigma^2_{\text{pop},1}, \ldots, \sigma^2_{\text{pop},K}$), $3 \times K$ at genotype level ($\sigma^2_{\text{gen},p,0}, \sigma^2_{\text{gen},p,1}, \sigma^2_{\text{gen},p}$ with $p = 1, \ldots, K$), $3 \times L$ at plant level ($\sigma^2_{\text{plant},g,0}, \sigma^2_{\text{plant},g,1}, \sigma^2_{\text{plant},g}$ with $g = 1, \ldots, L$), and the error variance (σ^2), while the number of regression coefficients (fixed and random) remains the same. It is worth noting that the implementation of the model with this specification allows for both genetic and plant-to-plant variations, or only one of them.

4.2.4 Covariance structure

Under the model specification (4.13) and based on the hierarchical structure in the psHDM (4.6),

$$\text{Cov}(\tilde{\boldsymbol{y}}) = \boldsymbol{Z}\text{Cov}(\boldsymbol{u})\boldsymbol{Z}^T + \boldsymbol{R} = \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^T + \boldsymbol{R}.$$

More specifically, there is the assumption that observations arising from the same plant, genotype or population are serially correlated, and the correlation increases as a function of the shared grouping levels (Brumback & Rice, 1998). In particular (and for simplicity), if we use the variance-covariance specification in (4.14), curves $\tilde{\mathbf{y}}_i$ and $\tilde{\mathbf{y}}_{i'}$ have covariance

$$cov(\tilde{\mathbf{y}}_i, \tilde{\mathbf{y}}_{i'}) = \begin{cases} \mathbf{0} & p(i) \neq p(i') \\ \boldsymbol{Z}_{\text{pop}}\boldsymbol{\Sigma}_{\text{pop}}\boldsymbol{Z}_{\text{pop}}^T & p(i) = p(i'), g(i) \neq g(i') \\ \boldsymbol{Z}_{\text{pop}}\boldsymbol{\Sigma}_{\text{pop}}\boldsymbol{Z}_{\text{pop}}^T + \boldsymbol{X}_{\text{gen}}\boldsymbol{\Sigma}_p^{\text{gen}}\boldsymbol{X}_{\text{gen}}^T + \sigma^2_{\text{gen},p}\boldsymbol{Z}_{\text{gen}}\boldsymbol{\Sigma}_{\text{gen}+}\boldsymbol{Z}_{\text{gen}}^T & g(i) = g(i'), i \neq i' \\ \boldsymbol{Z}_{\text{pop}}\boldsymbol{\Sigma}_{\text{pop}}\boldsymbol{Z}_{\text{pop}}^T + \boldsymbol{X}_{\text{gen}}\boldsymbol{\Sigma}_p^{\text{gen}}\boldsymbol{X}_{\text{gen}}^T + \sigma^2_{\text{gen},p}\boldsymbol{Z}_{\text{gen}}\boldsymbol{\Sigma}_{\text{gen}+}\boldsymbol{Z}_{\text{gen}}^T & \\ \quad + \boldsymbol{X}_{\text{plant}}\boldsymbol{\Sigma}_g^{\text{plant}}\boldsymbol{X}_{\text{plant}}^T + \sigma^2_{\text{plant},g}\boldsymbol{Z}_{\text{plant}}\boldsymbol{\Sigma}_{\text{plant}+}\boldsymbol{Z}_{\text{plant}}^T + \sigma^2\text{diag}(\boldsymbol{w}_i(t)) & i = i'. \end{cases}$$

This covariance structure is depicted in Figure 4.1 for a toy example with $K = 2$ populations, $L = 4$ genotypes, and $M = 8$ plants measured at $n = 10$ time points (a total of $8\times10 = 80$ observations). To observe differences between the two populations we deliberately increase the genetic variation and the variance of one of the populations ($p = 2$). We used cubic B-spline basis of dimension 13 to represent f_p, f_g and f_i (i.e., $b_{\text{pop}} = b_{\text{gen}} = b_{\text{plant}} = 13$). Under this configuration, the mixed model (4.13) has $2\times13+4\times13+8\times13 = 182$ regression coefficients (both fixed and random) and $(2 + 3 + 3 + 1=)$ 9 variance components.

This covariance structure is consistent with the empirical covariance structure of different datasets we have analysed before (see, e.g., Figure 2.8), as well as with other covariance structures reported in the literature for this kind of HTP data (e.g., Zhang, 2019). Figure 4.1 shows that for each plant trajectory the correlation increases with time, and the covariance increases as a function of the shared grouping levels. For instance, for population 2 ($p(i) = 2$) we observe some relation between trajectories of plants ($i = 5, \ldots, 8$) belonging to different genotypes ($g(i) = 3, 4$), but this relation increases for plant trajectories belonging to the same genotype (e.g., plants $i = 5, 6$ in genotype $g(i) = 3$).

4.2.5 psHDM estimation and computational aspects

We note that model in equation (4.13) is a standard linear mixed model. That is, the variance-covariance matrix for random effects in (4.14) has a standard form, with $\boldsymbol{G}$ linear in the variance parameters. Thus, estimation can be carried out with any mixed-model software (see, e.g., Bates et al., 2015; Butler et al., 2018; Pinheiro et al., 2019; SAS Institute Inc. 2015. SAS/STAT®, 2015; Wood, 2017). However, HTP data are usually characterised by a large number of observations, which, together with the number of regression

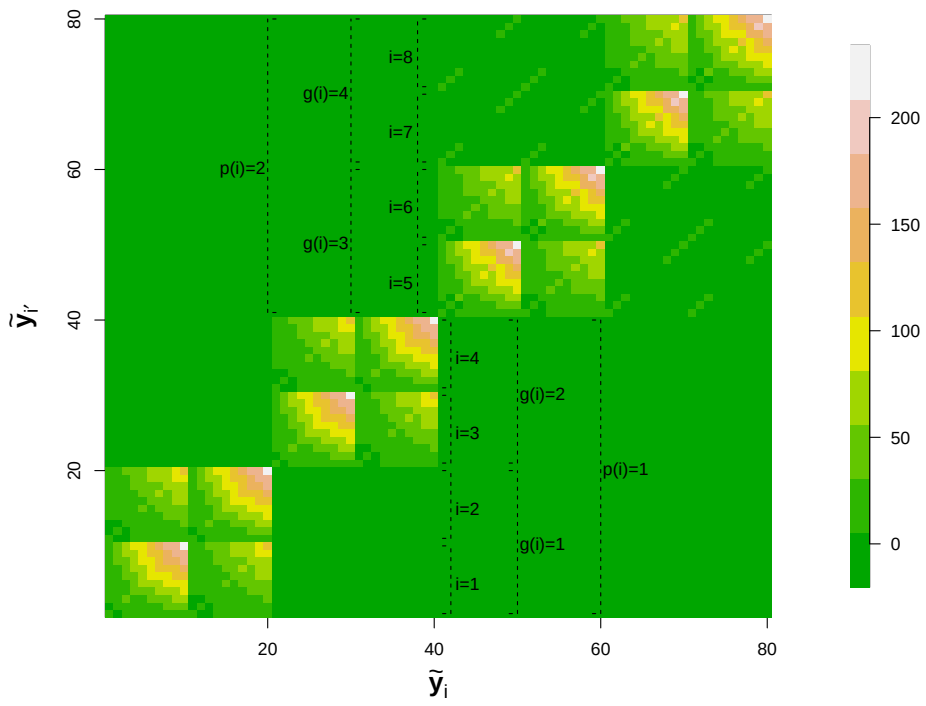


Figure 4.1: Variance-covariance structure under the model specification (4.13) for a toy example that considers $K = 2$ populations with $\sigma^2_{\text{pop},1} = 1$ and $\sigma^2_{\text{pop},2} = 5$, $L = 4$ genotypes with $\sigma^2_{\text{gen},p,0} = \sigma^2_{\text{gen},p,1} = 1$ and $\sigma^2_{\text{gen},p} = 10$, $M = 8$ plants with $\sigma^2_{\text{plant},g,0} = \sigma^2_{\text{plant},g,1} = \sigma^2_{\text{plant},g} = 1$, and $n = 10$ time points. Cubic B-spline bases of dimension 13 are used for the three levels of the hierarchy (i.e., $b_{\text{pop}} = b_{\text{gen}} = b_{\text{plant}} = 13$). Every pixel represents the covariance between the ith plant at the time t, $\tilde{y}_i(t)$, and the i'th plant at the time t, $\tilde{y}_{i'}(t)$.

coefficients (given by the selection of the cubic B-spline basis dimension) in equation (4.13), might make estimation with the above-mentioned software computationally expensive. Thus, we have implemented in the `R` language (R Core Team, 2023) our own code (freely available on https://CRAN.R-project.org/package=statgenHTP, Millet et al., 2022, through the function `fitSplineHDM()`), which resorts to the recently proposed SOP estimating algorithm, previously described in Section 3.3. However, as already mentioned, the variance-covariance matrix for the random effects in model (4.13) is linear in the variance parameters and thus, the SOP method reduces to the estimating algorithm described in Harville (1977).

BLUEs and BLUPs are obtained by the solution of Henderson's mixed model equations in Step 1 of the SOP algorithm. To speed up computation, we take advantage of the array structure of the data, which leads to the Kronecker structure of $\boldsymbol{X}$ and $\boldsymbol{Z}$ in (4.13), through the use of Generalised Linear Array Models (GLAM) (Currie et al., 2006). Specifically, we use the GLAM algorithm for fast and efficient computation of the matrix cross-products in $\boldsymbol{C}$ (see equation (3.35)). We also improve our codes by using efficient sparse matrix algebra implemented in the `spam` (Furrer et al., 2022) and `Matrix` (Bates et al., 2022) `R`-packages to construct the sparse matrices involved in the model ($\boldsymbol{X}$, $\boldsymbol{Z}$, $\boldsymbol{R}^{-1}$, $\boldsymbol{G}^{-1}$ and $\boldsymbol{C}$). Moreover, the SVD we use

to reformulate the P-spline model into the mixed model results in a diagonal variance-covariance matrix $\boldsymbol{G}$ (see (4.14)), making computation more efficient. One problem when solving the system of equations in (3.35) is that calculating $\boldsymbol{C}^{-1}$ is very expensive. We propose reducing time by using the sparse Cholesky decomposition of $\boldsymbol{C} = \boldsymbol{L}\boldsymbol{L}^T$, where $\boldsymbol{L}$ is a lower triangular matrix. Then, BLUEs and BLUPs are obtained by solving $\boldsymbol{L}\boldsymbol{L}^T\boldsymbol{b} = \tilde{\boldsymbol{y}}$ in (3.35) with $\boldsymbol{b} = (\hat{\boldsymbol{\beta}}^T, \hat{\boldsymbol{u}}^T)^T$ in two steps: (1) $\boldsymbol{L}\boldsymbol{z} = \boldsymbol{b}$ (forwards), and (2) $\boldsymbol{L}^T\tilde{\boldsymbol{y}} = \boldsymbol{z}$ (backwards).

Variance components are estimated using REML in Step 2 of the SOP algorithm. Here, we use the R-package `LMMsolver` (Boer, accepted, 2023; Boer & van Rossum, 2022) to calculate partial derivatives of the REML log-likelihood in an efficient way. The use of this package allows us to further reduce the computational burden by exploiting the sparse structure of the matrices involved in the model using the so-called 'sparse inverse', i.e., the automated differentiation of Cholesky algorithm proposed by S. P. Smith (1995). Specifically, the effective dimensions in (3.36) are equivalent to the partial derivatives of the log-determinant

$$\text{ED}_j^{[it]} = \hat{\sigma}_j^{-2[it]} \frac{\partial log \left|\boldsymbol{C}^{[it]}\right|}{\partial \hat{\sigma}_j^{-2[it]}}.$$

In this case, we do not need to calculate $\boldsymbol{C}^{-1}$, but its partial derivatives, which is more efficient.

4.2.6 Derivatives, standard errors and pointwise confidence intervals

As result of the second stage, we obtain estimated curves at the three levels of the hierarchy:

1. population trajectories ($\hat{f}_p$) and respective first-order derivatives ($\hat{f}'_p$),
2. genotype-specific deviations and first-order derivatives ($\hat{f}_g$, $\hat{f}'_g$), and respective trajectories and first-order derivatives ($\hat{f}_p + \hat{f}_g$, $(\hat{f}_p + \hat{f}_g)'$), and
3. plant-specific deviations and first-order derivatives ($\hat{f}_i$, $\hat{f}'_i$), respective trajectories and first-order derivatives ($\hat{f}_p + \hat{f}_g + \hat{f}_i$, $(\hat{f}_p + \hat{f}_g + \hat{f}_i)'$).

Derivatives are obtained as explained in Section 3.1.1.1. Construction of (approximate) confidence intervals for the estimated curves and their derivatives are based on the prediction error variance (for details go to Section 3.4 and Ruppert et al., 2003; Welham et al., 2004). Standard errors are based on $\boldsymbol{C}^{-1}$, thus two computational challenges arise at this point (in terms of large time and memory consumption): (i) calculation of $\boldsymbol{C}^{-1}$ (depending on the number of coefficients, i.e., the number of B-spline basis functions used) and, (ii) construction of confidence intervals for the estimated curves at the plant level (depending on the number of observations). For the former, we calculate the inverse of the variance-covariance matrix only once, at the last iteration step of the SOP algorithm. Regarding the second challenge, it is actually a limitation of our implementation and the user has to be aware that standard errors at the plant level will

demand large memory. Predictions are implemented in the R-function `predict.psHDM` of the `statgenHTP` package. They can be obtained for the same time points at which the original measurements were taken, or on a finer grid.

4.2.7 Extracting time-independent attributes to characterise genotypes

The second stage consists of a temporal analysis with a hierarchical curve data model to jointly estimate curves at each hierarchy level (plant or plot, genotype, and population) and their first-order derivatives. Different time-independent characteristics (intermediate traits) can be easily extracted from the estimated curves and their derivatives. Examples of intermediate traits for a growth-related trait (as illustrated in Figure 4.2, see, e.g., Hurtado et al., 2012; Roth et al., 2021) include the maximum and minimum trait values (from trajectories), the maximum and average growth rate (from the first-order derivative of the trajectories), the onset and end of senescence/growth (from second-order derivatives of the trajectories), and the area under the curve (AUC, from the deviation curves). Note that the timing of key plant-development stages can also be of interest. That is the value of the intermediate trait and the time at which it occurs. The area under the deviation curves can be interpreted as a global measure of a genotype/plant performance over time when compared to the genotypes/plants of the same region. A positive (negative) AUC indicates a genotype/plant performance better (worse) than the population/genotype average. The AUC can be estimated for the complete time interval where the measurements were taken. However, nothing precludes focusing attention on a restricted time interval of interest. Finally, if any treatment is applied to the trait of interest at any stage of the experiment, other maxima of the first-order derivative can be important to detect the response to the treatment (e.g., recovery rates). We will use these intermediate traits to compare the one and two-stage approaches proposed in this thesis in Chapters 6 and 7.

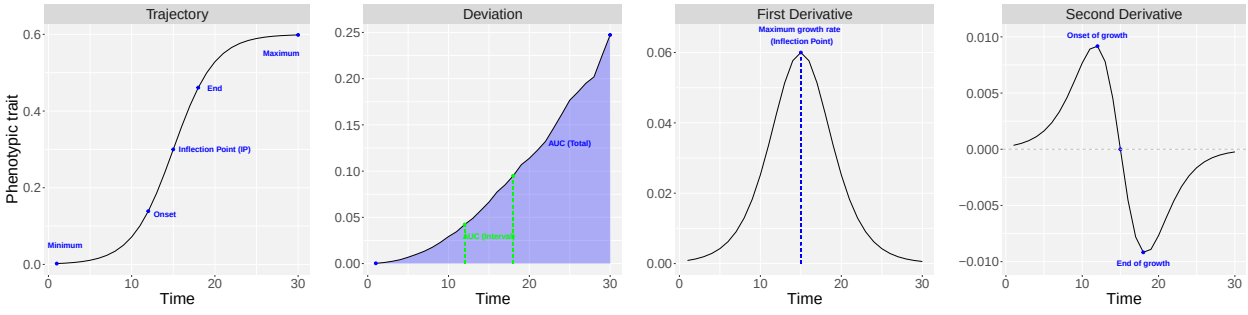


Figure 4.2: Examples of different intermediate traits (time-independent characteristics) obtained from estimated curves at each level of the hierarchy (plant or plot, genotype, and population) as well as their first-order derivatives. We use a growth-related trait as illustration.

Chapter 5

Spatio-temporal modelling of high-throughput phenotyping data: One-stage approach

In the previous chapter, we introduced a two-stage approach to model hierarchical spatio-temporal data from HTP experiments. Although stage-wise proposals are computationally feasible, they may result in loss of information between and within stages. For instance, our two-stage P-splines-based approach may not fully account for spatial heterogeneity across time when correcting for environmental factors in the first stage, and uncertainty is lost between stages (weights are used to propagate error from the first to the second stage). It is therefore of interest to develop approaches that allow modelling the spatial and temporal genetic and non-genetic variation in one stage to take advantage of all the available information. To that aim, in this chapter, we propose a one-stage spatio-temporal P-spline hierarchical curve data model for the analysis of HTP data. In particular, we generalise the two-stage modelling strategy presented in Chapter 4 to a full and one-stage spatio-temporal approach. We use the SpATS model as the base model and extend it to the spatio-temporal case by considering a three-level hierarchical data structure (populations, genotypes within populations, and plants within genotypes) and a three-dimensional smooth function (similar to the f_{ST} presented in Sections 3.1.2.2 and 3.2.3). While the transition from a two-stage to a one-stage approach may seem straightforward, implementing the latter is challenging due to the complexity and dimensionality of the data and models involved, leading to issues such as identifiability, scalability, and computational burden. We assess the performance of the proposed approach using simulated and real data in Chapters 6 and 7, and compare the results with those obtained using the two-stage approach. Additionally, we present advances in software implementation in Chapter 8. This chapter builds on the material presented in Perez-Valencia et al. (2023) and represents a significant step towards a more comprehensive and efficient solution for the analysis of HTP data.

5.1 Spatio-temporal (psHDM) P-spline hierarchical curve data model

Our approach builds upon the spatial SpATS model proposed by Rodríguez-Álvarez et al. (2018). In particular, if y_i is the phenotypic trait for the ith plant for a specific time point t (for simplicity, we omit here the dependence on time), the SpATS model we consider is as follows

$$y_i = f_{p(i)} + f_{g(i)} + f_{r(i)} + f_{c(i)} + \underbrace{f_{\mathrm{S}}(r(i), c(i))}_{\text{Spatial trend}} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2), \quad 1 \leq i \leq M. \tag{5.1}$$

Note that model (5.1) is a simpler model than the one presented in the first stage of the two-stage approach (4.1); here we do not consider extra experimental design factors. Recall that f_p is the fixed effect coefficient for population p, f_g is the random effect coefficient for genotype g ($f_g \sim N(0, \sigma^2_{\text{gen}})$), and f_r and f_c are random effect coefficients for row r and column c, respectively ($f_r \sim N(0, \sigma^2_{\text{row}})$ and $f_c \sim N(0, \sigma^2_{\text{col}})$). Finally, $f_{\mathrm{S}}(r, c)$ is a two-dimensional smooth function at time t, defined over the row and column positions, that simultaneously accounts for the spatial (local and global) trend variation across both directions. This smooth function is constructed with tensor-product P-splines (see equation (3.20), and Eilers & Marx, 1996, 2003).

By taking the SpATS model (5.1) as the base model, we now extend it to the spatio-temporal case by allowing all effects to vary with time, i.e.,

$$\begin{aligned} y_i(t) &= f_{p(i)}(t) + f_{g(i)}(t) + f_{r(i)}(t) + f_{c(i)}(t) + f_{\mathrm{ST}}(r(i), c(i), t) + \varepsilon_i(t), \\ &= \underbrace{f_{p(i)}(t) + f_{g(i)}(t) + f_i(t)}_{\text{3-level longitudinal effects}} + f_{r(i)}(t) + f_{c(i)}(t) + \underbrace{f_{\mathrm{ST}}(r(i), c(i), t)}_{\text{Spatio-temporal trend}} + \varepsilon_i. \end{aligned} \tag{5.2}$$

Note that, in the second equation, $\varepsilon_i(t)$ has been decomposed such that $\varepsilon_i(t) = f_i(t) + \varepsilon_i$, that is, we capture the temporal trend of each plant in $f_i(t)$, while ε_i is pure random noise, i.e., $\varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$. The interpretation of each component in model (5.2) is as follows: $f_p(t)$ is the time-varying effect coefficient for population p; $f_g(t)$ is the time-varying random effect coefficient for genotype g (it measures deviations from the population effect to which the genotype belongs to); $f_i(t)$ is the time-varying random effect coefficient for plant i (it measures deviations from the genotype effect to which the plant belongs to); and $f_r(t)$ and $f_c(t)$ are time-varying random effect coefficients for row r and column c, respectively. Finally, $f_{\mathrm{ST}}(r, c, t)$ is a spatio-temporal three-dimensional surface defined over rows, columns and time. This three-dimensional surface accounts for spatial trend variations, but it allows these spatial trends to change with time.

To model (and estimate) the time-varying effects in model (5.2), we assume that all these effects vary smoothly along time. Following the ideas presented in Section 3.1.1, each one-dimensional function in (5.2), $f_p(t)$, $f_g(t)$, $f_i(t)$, $f_r(t)$, and $f_c(t)$, is modelled as a linear combination of cubic B-spline basis functions. Similarly, in Section 3.1.2.2, we presented the basis for modeling the three-dimensional function $f_{\mathrm{ST}}(r, c, t)$ using the tensor product of three marginal cubic B-spline bases. From now on we call model (5.2) the

spatio-temporal psHDM, which is approximated by a linear combination of cubic B-spline basis functions as follows

$$\begin{aligned} y_i(t) = &\underbrace{\sum_{k_{\text{pop}}=1}^{b_{\text{pop}}} B_{k_{\text{pop}}}(t)\theta^{\text{pop}}_{p(i),k_{\text{pop}}}}_{f_{p(i)}(t)} + \underbrace{\sum_{k_{\text{gen}}=1}^{b_{\text{gen}}} B_{k_{\text{gen}}}(t)\theta^{\text{gen}}_{g(i),k_{\text{gen}}}}_{f_{g(i)}(t)} + \underbrace{\sum_{k_{\text{plant}}=1}^{b_{\text{plant}}} B_{k_{\text{plant}}}(t)\theta^{\text{plant}}_{i,k_{\text{plant}}}}_{f_i(t)} \\ &+ \underbrace{\sum_{k_{\text{row}}=1}^{b_{\text{row}}} B_{k_{\text{row}}}(t)\theta^{\text{row}}_{r(i),k_{\text{row}}}}_{f_{r(i)}(t)} + \underbrace{\sum_{k_{\text{col}}=1}^{b_{\text{col}}} B_{k_{\text{col}}}(t)\theta^{\text{col}}_{c(i),k_{\text{col}}}}_{f_{c(i)}(t)} + f_{\text{ST}}(r(i), c(i), t) + \varepsilon_i(t), \end{aligned} \tag{5.3}$$

where $(\boldsymbol{B}_{\text{pop}})^{n\times b_{\text{pop}}}_{jk_{\text{pop}}} = B_{k_{\text{pop}}}(t_j)$, $(\boldsymbol{B}_{\text{gen}})^{n\times b_{\text{gen}}}_{jk_{\text{gen}}} = B_{k_{\text{gen}}}(t_j)$ and $(\boldsymbol{B}_{\text{plant}})^{n\times b_{\text{plant}}}_{jk_{\text{plant}}} = B_{k_{\text{plant}}}(t_j)$ are cubic B-spline basis functions, evaluated at time t_j, at population, genotype and plant levels, respectively; $(\boldsymbol{B}_{\text{row}})^{n\times b_{\text{row}}}_{jk_{\text{row}}} = B_{k_{\text{row}}}(t_j)$ and $(\boldsymbol{B}_{\text{col}})^{n\times b_{\text{col}}}_{jk_{\text{col}}} = B_{k_{\text{col}}}(t_j)$ are cubic B-spline basis functions, evaluated at time t_j, for rows and columns, respectively; and f_{ST} is the tensor-product of three one-dimensional cubic B-splines basis (in the row, column and time directions) in (3.13), which results in the spatio-temporal B-spline design matrix $\boldsymbol{B}_{\text{ST}}$ in (3.14). Here, $\boldsymbol{\theta}^{\text{pop}}_{p(i)} = (\theta^{\text{pop}}_{p(i),1}, \ldots, \theta^{\text{pop}}_{p(i),b_{\text{pop}}})^T$, $\boldsymbol{\theta}^{\text{gen}}_{g(i)} = (\theta^{\text{gen}}_{g(i),1}, \ldots, \theta^{\text{gen}}_{g(i),b_{\text{gen}}})^T$, $\boldsymbol{\theta}^{\text{plant}}_{i} = (\theta^{\text{plant}}_{i,1}, \ldots, \theta^{\text{plant}}_{i,b_{\text{plant}}})^T$, $\boldsymbol{\theta}^{\text{row}}_{r(i)} = (\theta^{\text{row}}_{r(i),1}, \ldots, \theta^{\text{row}}_{r(i),b_{\text{row}}})^T$, and $\boldsymbol{\theta}^{\text{col}}_{c(i)} = (\theta^{\text{col}}_{c(i),1}, \ldots, \theta^{\text{col}}_{c(i),b_{\text{col}}})^T$ are vectors of unknown regression coefficients.

We now present the P-spline model (5.3) in matrix notation for all plants $\boldsymbol{y} = (y_1(t_1), \ldots, y_1(t_n), \ldots, y_M(t_1), \ldots, y_M(t_n))^T$, which are ordered by plant and time. As for the two-stage approach, recall that data are pre-ordered by population, genotype, plant and time to make use of the Kronecker products

$$\begin{aligned} \boldsymbol{y} = &\underbrace{(\boldsymbol{Q}_{\text{pop}} \otimes \boldsymbol{B}_{\text{pop}})\boldsymbol{\theta}_{\text{pop}}}_{\boldsymbol{f}_{\text{pop}}} + \underbrace{(\boldsymbol{Q}_{\text{gen}} \otimes \boldsymbol{B}_{\text{gen}})\boldsymbol{\theta}_{\text{gen}}}_{\boldsymbol{f}_{\text{gen}}} + \underbrace{(\boldsymbol{I}_M \otimes \boldsymbol{B}_{\text{plant}})\boldsymbol{\theta}_{\text{plant}}}_{\boldsymbol{f}_{\text{plant}}} \\ &+ \underbrace{(\boldsymbol{Q}_{\text{row}} \otimes \boldsymbol{B}_{\text{row}})\boldsymbol{\theta}_{\text{row}}}_{\boldsymbol{f}_{\text{row}}} + \underbrace{(\boldsymbol{Q}_{\text{col}} \otimes \boldsymbol{B}_{\text{col}})\boldsymbol{\theta}_{\text{col}}}_{\boldsymbol{f}_{\text{col}}} + \underbrace{\boldsymbol{B}_{\text{ST}}\boldsymbol{\theta}_{\text{ST}}}_{\boldsymbol{f}_{\text{ST}}} + \boldsymbol{\varepsilon}, \end{aligned} \tag{5.4}$$

where $\boldsymbol{f}_{\text{pop}} = (f_{p(1)}(t_1), \ldots, f_{p(1)}(t_n), \ldots, f_{p(M)}(t_1), \ldots, f_{p(M)}(t_n))^T$, $\boldsymbol{f}_{\text{gen}} = (f_{g(1)}(t_1), \ldots, f_{g(1)}(t_n), \ldots, f_{g(M)}(t_1), \ldots, f_{g(M)}(t_n))^T$, $\boldsymbol{f}_{\text{plant}} = (f_1(t_1), \ldots, f_1(t_n), \ldots, f_M(t_1), \ldots, f_M(t_n))^T$, $\boldsymbol{f}_{\text{row}} = (f_{r(1)}(t_1), \ldots, f_{r(1)}(t_n), \ldots, f_{r(M)}(t_1), \ldots, f_{r(M)}(t_n))^T$, $\boldsymbol{f}_{\text{col}} = (f_{c(1)}(t_1), \ldots, f_{c(1)}(t_n), \ldots, f_{c(M)}(t_1), \ldots, f_{c(M)}(t_n))^T$, and $\boldsymbol{f}_{\text{ST}}$ as defined in (3.13). Additionally, $\boldsymbol{Q}_{\text{pop}}$, $\boldsymbol{Q}_{\text{gen}}$, $\boldsymbol{Q}_{\text{row}}$ and $\boldsymbol{Q}_{\text{col}}$ are contrast matrices assigning, respectively, plants to populations, plants to genotypes, plants to row locations, and plants to column locations. That is, $\boldsymbol{Q}^{M\times K}_{\text{pop}} = \text{blockdiag}(\mathbf{1}^{\text{pop},T}_1, \ldots, \mathbf{1}^{\text{pop},T}_K)$, with $\mathbf{1}^{\text{pop}}_p$ vectors of ones of length $\#\{i \mid p(i) = p\}$; $\boldsymbol{Q}^{M\times L}_{\text{gen}} = \text{blockdiag}(\mathbf{1}^{\text{gen},T}_1, \ldots, \mathbf{1}^{\text{gen},T}_L)$, with $\mathbf{1}^{\text{gen}}_g$ vectors of ones of length m_g; $\boldsymbol{Q}^{M\times R}_{\text{row}} = \text{blockdiag}(\mathbf{1}^{\text{row},T}_1, \ldots, \mathbf{1}^{\text{row},T}_R)$, with $\mathbf{1}^{\text{row}}_r$ vectors of ones of length $\#\{i \mid r(i) = r\}$; and $\boldsymbol{Q}^{M\times C}_{\text{col}} = \text{blockdiag}(\mathbf{1}^{\text{col},T}_1, \ldots, \mathbf{1}^{\text{col},T}_C)$, with $\mathbf{1}_{\text{col},c}$ vectors of ones of length $\#\{i \mid c(i) = c\}$. The vectors of unknown regression coefficients for all plants become $\boldsymbol{\theta}_{\text{pop}} = (\boldsymbol{\theta}^{\text{pop};T}_1, \ldots, \boldsymbol{\theta}^{\text{pop};T}_K)^T$, $\boldsymbol{\theta}_{\text{gen}} = (\boldsymbol{\theta}^{\text{gen};T}_1, \ldots, \boldsymbol{\theta}^{\text{gen};T}_L)^T$, $\boldsymbol{\theta}_{\text{plant}} = (\boldsymbol{\theta}^{\text{plant};T}_1, \ldots, \boldsymbol{\theta}^{\text{plant};T}_M)^T$, $\boldsymbol{\theta}_{\text{row}} = (\boldsymbol{\theta}^{\text{row};T}_1, \ldots, \boldsymbol{\theta}^{\text{row};T}_R)^T$, and $\boldsymbol{\theta}_{\text{col}} = (\boldsymbol{\theta}^{\text{col};T}_1, \ldots, \boldsymbol{\theta}^{\text{col};T}_C)^T$ (see (3.13) for $\boldsymbol{\theta}_{\text{ST}}$).

As said before, we propose using P-splines (Eilers & Marx, 1996, 2003) to model $\boldsymbol{y}$, i.e., we combine B-spline basis functions on equidistant knots and a discrete difference penalty on the (regression) coefficients to ensure smoothness. Then, the penalty associated with model (5.4) is

$$\begin{aligned}\sum_{p=1}^{K}\lambda_{\text{pop},p}\boldsymbol{\theta}_p^{\text{pop};T}\boldsymbol{P}_{\text{pop}}\boldsymbol{\theta}_p^{\text{pop}} + \lambda_{\text{gen}}\boldsymbol{\theta}_{\text{gen}}^T(\boldsymbol{I}_L\otimes\boldsymbol{P}_{\text{gen}})\boldsymbol{\theta}_{\text{gen}} + \lambda_{\text{plant}}\boldsymbol{\theta}_{\text{plant}}^T(\boldsymbol{I}_M\otimes\boldsymbol{P}_{\text{plant}})\boldsymbol{\theta}_{\text{plant}} + \\ \lambda_{\text{row}}\boldsymbol{\theta}_{\text{row}}^T(\boldsymbol{I}_R\otimes\boldsymbol{P}_{\text{row}})\boldsymbol{\theta}_{\text{row}} + \lambda_{\text{col}}\boldsymbol{\theta}_{\text{col}}^T(\boldsymbol{I}_C\otimes\boldsymbol{P}_{\text{col}})\boldsymbol{\theta}_{\text{col}} + \boldsymbol{\theta}_{\text{ST}}^T\boldsymbol{P}_{\text{ST}}\boldsymbol{\theta}_{\text{ST}},\end{aligned} \tag{5.5}$$

where $\boldsymbol{P}_\nu = \boldsymbol{D}_\nu^T\boldsymbol{D}_\nu$ ($\nu \in \{\text{pop}, \text{gen}, \text{plant}, \text{row}, \text{col}\}$) are penalty matrices with $\boldsymbol{D}_\nu$ matrices that form second order differences, and $\boldsymbol{P}_{\text{ST}}$ as defined in (3.15). The influence of the penalty (i.e., the amount of smoothness) is determined by one smoothing parameter for each one-dimensional function (i.e., λ_{gen}, λ_{plant}, λ_{row} and λ_{col}), and by three smoothing parameters in the three-dimensional case (with λ_1, λ_2 and λ_3 smoothing parameters in the time, row and column directions), i.e., we consider anisotropy (details are given in Sections 3.2.1 and 3.2.3, but we refer the reader to Eilers & Marx, 2021; Rodríguez-Álvarez et al., 2015, and references therein for a more in-depth presentation). As for the second stage of our two-stage approach, we allow for different population smoothing parameters (i.e., $\boldsymbol{\lambda}_{\text{pop}} = (\lambda_{\text{pop},1}, \ldots, \lambda_{\text{pop},K})$). In what follows, we present in detail the estimation of the spatio-temporal psHDM (5.2) and discuss some computational aspects.

5.2 Mixed model formulation of the spatio-temporal psHDM

For estimation, we follow the same modeling philosophy used for the two-stage approach. That is, we use the connection between P-splines and linear mixed models (Currie & Durban, 2002; Currie et al., 2006; Lee & Durban, 2011; Wand, 2003). Thus, the smooth functions are decomposed (reparameterised) in two components: one whose coefficients are not penalised (these coefficients are considered as fixed effect coefficients), and one whose coefficients are penalised (and thus considered as random effect coefficients). As before, smoothing parameters become (ratios of) variance parameters, and estimated using REML. Following the same ideas as for the two-stage approach, the unpenalised/fixed effect coefficients (intercept and slope) for $f_g(t)$ and $f_i(t)$ are assumed to be penalised/random to avoid identifiability problems (and the same assumption applies to $f_r(t)$ and $f_c(t)$). In summary, the modeling strategy we follow implies that there is one variance parameter per population, and three variance parameters (associated, respectively, with the intercept, the slope and the non-linear/penalised/smooth effect) for genotypes, plants, rows and columns. Thus, under this framework, the spatio-temporal psHDM (5.2) is expressed as follows (for one plant) (we refer the

reader to Section 3.2.1 for more technical details on the spatio-temporal smooth function, f_{ST})

$$\begin{aligned}\boldsymbol{y}_i = &\underbrace{\boldsymbol{X}_{\text{pop}}\boldsymbol{\beta}_{\text{lin},p(i)} + \boldsymbol{Z}_{\text{pop}}\boldsymbol{u}^{\text{pop}}_{\text{nlin},p}}_{f_p} + \underbrace{\boldsymbol{X}_{\text{gen}}\boldsymbol{u}^{\text{gen}}_{\text{lin},g(i)} + \boldsymbol{Z}_{\text{gen}}\boldsymbol{u}^{\text{gen}}_{\text{nlin},g(i)}}_{f_g} + \underbrace{\boldsymbol{X}_{\text{plant}}\boldsymbol{u}^{\text{plant}}_{\text{lin},i} + \boldsymbol{Z}_{\text{plant}}\boldsymbol{u}^{\text{plant}}_{\text{nlin},i}}_{f_i} + \\ &\underbrace{\boldsymbol{X}_{\text{row}}\boldsymbol{u}^{\text{row}}_{\text{lin},r(i)} + \boldsymbol{Z}_{\text{row}}\boldsymbol{u}^{\text{row}}_{\text{nlin},r(i)}}_{f_r} + \underbrace{\boldsymbol{X}_{\text{col}}\boldsymbol{u}^{\text{col}}_{\text{lin},c(i)} + \boldsymbol{Z}_{\text{col}}\boldsymbol{u}^{\text{col}}_{\text{nlin},c(i)}}_{f_c} + \boldsymbol{f}_{\text{ST}}(r(i), c(i), \boldsymbol{t}) + \boldsymbol{\varepsilon}_i,\end{aligned} \tag{5.6}$$

where the basis functions, the regression coefficients, and the variance-covariance matrices of the random effects, are obtained from the connection between P-splines and linear mixed models (see Section 3.2.1). As before, we use the SVD of the penalty matrices $\boldsymbol{P}_\nu = \boldsymbol{D}_\nu^T\boldsymbol{D}_\nu = \boldsymbol{U}_\nu\boldsymbol{\Lambda}_\nu\boldsymbol{U}_\nu^T$ ($\nu \in \{\text{pop}, \text{gen}, \text{plant}, \text{row}, \text{col}\}$) to find these expressions. Therefore, the specification of the model (5.6) is as follows

$$\begin{aligned}
\boldsymbol{X}_{\text{pop}} &= [\boldsymbol{1}_n \mid \boldsymbol{t}] \text{ and } \boldsymbol{\beta}^{\text{pop}}_{\text{lin},p} = (\beta_{\text{lin},p,0}, \beta_{\text{lin},p,1})^T,\\
\boldsymbol{Z}_{\text{pop}} &= \boldsymbol{B}_{\text{pop}}\boldsymbol{U}_{\text{pop}+} \text{ and } \boldsymbol{u}^{\text{pop}}_{\text{nlin},p} = \boldsymbol{U}^T_{\text{pop}+}\boldsymbol{\theta}^{\text{pop}}_p = (u^{\text{pop}}_{\text{nlin},p,1}, \ldots, u^{\text{pop}}_{\text{nlin},p,b_{\text{pop}}-2})^T \sim N(\boldsymbol{0}, \sigma^2_{\text{pop},p}\boldsymbol{\Sigma}_{\text{pop}+}),\\
\boldsymbol{X}_{\text{gen}} &= [\boldsymbol{1}_n \mid \boldsymbol{t}] \text{ and } \boldsymbol{u}^{\text{gen}}_{\text{lin},g} = (u^{\text{gen}}_{\text{lin},g,0}, u^{\text{gen}}_{\text{lin},g,1})^T \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_{\text{gen}}) \text{ with } \boldsymbol{\Sigma}_{\text{gen}} = \text{diag}(\sigma^2_{\text{gen},0}, \sigma^2_{\text{gen},1}),\\
\boldsymbol{Z}_{\text{gen}} &= \boldsymbol{B}_{\text{gen}}\boldsymbol{U}_{\text{gen}+} \text{ and } \boldsymbol{u}^{\text{gen}}_{\text{nlin},g} = \boldsymbol{U}^T_{\text{gen}+}\boldsymbol{\theta}^{\text{gen}}_g = (u^{\text{gen}}_{\text{nlin},g,1}, \ldots, u^{\text{gen}}_{\text{nlin},g,b_{\text{gen}}-2})^T \sim N(\boldsymbol{0}, \sigma^2_{\text{gen}}\boldsymbol{\Sigma}_{\text{gen}+}),\\
\boldsymbol{X}_{\text{plant}} &= [\boldsymbol{1}_n \mid \boldsymbol{t}] \text{ and } \boldsymbol{u}^{\text{plant}}_{\text{lin},i} = (u^{\text{plant}}_{\text{lin},i,0}, u^{\text{plant}}_{\text{lin},i,1})^T \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_{\text{plant}}) \text{ with } \boldsymbol{\Sigma}_{\text{plant}} = \text{diag}(\sigma^2_{\text{plant},0}, \sigma^2_{\text{plant},1}),\\
\boldsymbol{Z}_{\text{plant}} &= \boldsymbol{B}_{\text{plant}}\boldsymbol{U}_{\text{plant}+} \text{ and } \boldsymbol{u}^{\text{plant}}_{\text{nlin},i} = \boldsymbol{U}^T_{\text{plant}+}\boldsymbol{\theta}^{\text{plant}}_i = (u^{\text{plant}}_{\text{nlin},i,1}, \ldots, u^{\text{plant}}_{\text{nlin},i,b_{\text{plant}}-2})^T \sim N(\boldsymbol{0}, \sigma^2_{\text{plant}}\boldsymbol{\Sigma}_{\text{plant}+}),\\
\boldsymbol{X}_{\text{row}} &= [\boldsymbol{1}_n \mid \boldsymbol{t}] \text{ and } \boldsymbol{u}^{\text{row}}_{\text{lin},r} = (u^{\text{row}}_{\text{lin},r,0}, u^{\text{row}}_{\text{lin},r,1})^T \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_{\text{row}}) \text{ with } \boldsymbol{\Sigma}_{\text{row}} = \text{diag}(\sigma^2_{\text{row},0}, \sigma^2_{\text{row},1}),\\
\boldsymbol{Z}_{\text{row}} &= \boldsymbol{B}_{\text{row}}\boldsymbol{U}_{\text{row}+} \text{ and } \boldsymbol{u}^{\text{row}}_{\text{nlin},r} = \boldsymbol{U}^T_{\text{row}+}\boldsymbol{\theta}^{\text{row}}_r = (u^{\text{row}}_{\text{nlin},r,1}, \ldots, u^{\text{row}}_{\text{nlin},r,b_{\text{row}}-2})^T \sim N(\boldsymbol{0}, \sigma^2_{\text{row}}\boldsymbol{\Sigma}_{\text{row}+}),\\
\boldsymbol{X}_{\text{col}} &= [\boldsymbol{1}_n \mid \boldsymbol{t}] \text{ and } \boldsymbol{u}^{\text{col}}_{\text{lin},c} = (u^{\text{col}}_{\text{lin},c,0}, u^{\text{col}}_{\text{lin},c,1})^T \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_{\text{col}}) \text{ with } \boldsymbol{\Sigma}_{\text{col}} = \text{diag}(\sigma^2_{\text{col},0}, \sigma^2_{\text{col},1}),\\
\boldsymbol{Z}_{\text{col}} &= \boldsymbol{B}_{\text{col}}\boldsymbol{U}_{\text{col}+} \text{ and } \boldsymbol{u}^{\text{col}}_{\text{nlin},c} = \boldsymbol{U}^T_{\text{col}+}\boldsymbol{\theta}^{\text{col}}_c = (u^{\text{col}}_{\text{nlin},c,1}, \ldots, u^{\text{col}}_{\text{nlin},c,b_{\text{col}}-2})^T \sim N(\boldsymbol{0}, \sigma^2_{\text{col}}\boldsymbol{\Sigma}_{\text{col}+}),
\end{aligned} \tag{5.7}$$

where the variance-covariance matrices for $\boldsymbol{u}^{\text{pop}}_{\text{nlin},p}$, $\boldsymbol{u}^{\text{gen}}_{\text{nlin},g}$, $\boldsymbol{u}^{\text{plant}}_{\text{nlin},i}$, $\boldsymbol{u}^{\text{row}}_{\text{nlin},r}$, and $\boldsymbol{u}^{\text{col}}_{\text{nlin},c}$ are obtained from the inverse of their respective precision matrices. That is, $\sigma^2_\nu\boldsymbol{\Sigma}_{\nu+} = \lambda^{-1}_\nu\boldsymbol{\Lambda}^{-1}_{\nu+}$, with $\sigma^2_\nu = \sigma^2/\lambda_\nu$ and $\boldsymbol{\Sigma}_{\nu+} = \boldsymbol{\Lambda}^{-1}_{\nu+}$ ($\nu \in \{\text{gen}, \text{plant}, \text{row}, \text{col}\}$), and $\sigma^2_{\text{pop},p}\boldsymbol{\Sigma}_{\text{pop}+} = \lambda^{-1}_{\text{pop},p}\boldsymbol{\Lambda}^{-1}_{\text{pop}+}$, with $\sigma^2_{\text{pop},p} = \sigma^2/\lambda_{\text{pop},p}$ and $\boldsymbol{\Sigma}_{\text{pop}+} = \boldsymbol{\Lambda}^{-1}_{\text{pop}+}$ at the population level (we consider different smoothing/variance parameters for each population). For simplicity, we use the parameterization $\boldsymbol{X}_\nu = [\boldsymbol{1}_n \mid \boldsymbol{t}]$ instead of $\boldsymbol{X}_\nu = \boldsymbol{B}_\nu\boldsymbol{U}_{\nu+}$, $\nu \in \{\text{pop}, \text{gen}, \text{plant}, \text{row}, \text{col}\}$, (but more precisely, they are obtained as described in Wood et al. (2013); see also Section 3.2.1 for more details), with $\boldsymbol{\beta}^\nu_{\text{lin},\hat{\nu}} = \boldsymbol{u}^\nu_{\text{lin},\hat{\nu}} = (u^\nu_{\text{lin},\hat{\nu},0}, u^\nu_{\text{lin},\hat{\nu},1})^T$, $\hat{\nu} \in \{g, i, r, c\}$, and $\boldsymbol{\beta}^{\text{pop}}_{\text{lin},p} = (\beta^{\text{pop}}_{\text{lin},p,0}, \beta^{\text{pop}}_{\text{lin},p,1})^T$ the corresponding intercepts and slopes. With respect to the spatio-temporal three-dimensional surface, $\boldsymbol{f}_{\text{ST}}(r, c, t)$, we follow the ideas in Lee and Durban (2011) and Rodríguez-Álvarez et al. (2015) for its mixed model reparameterisation (for details, see also Section 3.2.3).

Before proceeding, recall that $\boldsymbol{f}_{\text{ST}}$ can be decomposed in an ANOVA-way (see (3.34)). This decomposition reveals that there are three components which are confounded with the population effect, namely, the intercept β_0, the linear effect along time $t \times \beta_3$, and the non-linear (smooth) main effect along time $f_1(t)$.

These terms (in blue) are removed from f_{ST}

$$
\begin{aligned}
\boldsymbol{f}_{\text{ST}} = \underbrace{\mathbf{1}_{Mn}\beta_0 + (\boldsymbol{r}\otimes\mathbf{1}_n)\beta_1 + (\boldsymbol{c}\otimes\mathbf{1}_n)\beta_2 + (\boldsymbol{t}\otimes\mathbf{1}_M)\beta_3 + ((\boldsymbol{r}\odot\boldsymbol{c})\otimes\mathbf{1}_n)\beta_4 + (\boldsymbol{r}\otimes\boldsymbol{t})\beta_5 + (\boldsymbol{c}\otimes\boldsymbol{t})\beta_6 + ((\boldsymbol{r}\odot\boldsymbol{c})\otimes\boldsymbol{t})\beta_7}_{\text{Linear effects and interactions}} + \\
\underbrace{\underbrace{f_2(\boldsymbol{r})\otimes\mathbf{1}_n}_{(\mathbf{Z}_2\otimes\mathbf{1}_n)\boldsymbol{u}_2} + \underbrace{(\boldsymbol{c}\odot h_{2;3}(\boldsymbol{r}))\otimes\mathbf{1}_n}_{((\mathbf{Z}_2\square\boldsymbol{c})\otimes\mathbf{1}_n)\boldsymbol{u}_{2;3}} + \underbrace{h_{2;1}(\boldsymbol{r})\otimes\boldsymbol{t}}_{(\mathbf{Z}_2\otimes\boldsymbol{t})\boldsymbol{u}_{2;1}} + \underbrace{(\boldsymbol{c}\odot h_{2;31}(\boldsymbol{r}))\otimes\boldsymbol{t}}_{((\mathbf{Z}_2\square\boldsymbol{c})\otimes\boldsymbol{t})\boldsymbol{u}_{2;31}}}_{\text{Non-linear row-related effects}} + \\
\underbrace{\underbrace{f_3(\boldsymbol{c})\otimes\mathbf{1}_n}_{(\mathbf{Z}_3\otimes\mathbf{1}_n)\boldsymbol{u}_3} + \underbrace{(\boldsymbol{r}\odot h_{3;2}(\boldsymbol{c}))\otimes\mathbf{1}_n}_{((\boldsymbol{r}\square\mathbf{Z}_3)\otimes\mathbf{1}_n)\boldsymbol{u}_{3;2}} + \underbrace{h_{3;1}(\boldsymbol{c})\otimes\boldsymbol{t}}_{(\mathbf{Z}_3\otimes\boldsymbol{t})\boldsymbol{u}_{3;1}} + \underbrace{(\boldsymbol{r}\odot h_{3;21}(\boldsymbol{c}))\otimes\boldsymbol{t}}_{((\boldsymbol{r}\square\mathbf{Z}_3)\otimes\boldsymbol{t})\boldsymbol{u}_{3;21}}}_{\text{Non-linear column-related effects}} + \\
\underbrace{\underbrace{f_1(\boldsymbol{t})\otimes\mathbf{1}_M}_{(\mathbf{Z}_{\text{T}}\otimes\mathbf{1}_M)\boldsymbol{u}_1} + \underbrace{\boldsymbol{r}\otimes h_{1;2}(\boldsymbol{t})}_{(\boldsymbol{r}\otimes\mathbf{Z}_{\text{T}})\boldsymbol{u}_{1;2}} + \underbrace{\boldsymbol{c}\otimes h_{1;3}(\boldsymbol{t})}_{(\boldsymbol{c}\otimes\mathbf{Z}_{\text{T}})\boldsymbol{u}_{1;3}} + \underbrace{(\boldsymbol{r}\odot\boldsymbol{c})\otimes h_{1;23}(\boldsymbol{t})}_{((\boldsymbol{r}\odot\boldsymbol{c})\otimes\mathbf{Z}_{\text{T}})\boldsymbol{u}_{1;23}}}_{\text{Non-linear time-related effects}} + \\
\underbrace{\underbrace{h_{2|3}(\boldsymbol{r},\boldsymbol{c})\otimes\mathbf{1}_M}_{((\mathbf{Z}_2\square\mathbf{Z}_3)\otimes\mathbf{1}_M)\boldsymbol{u}_{2|3}} + \underbrace{h_{2|3;1}(\boldsymbol{r},\boldsymbol{c})\otimes\boldsymbol{t}}_{((\mathbf{Z}_2\square\mathbf{Z}_3)\otimes\boldsymbol{t})\boldsymbol{u}_{2|3;1}}}_{\text{Non-linear row and column interactions}} + \underbrace{\underbrace{h_{2|1}(\boldsymbol{r},\boldsymbol{t})}_{(\mathbf{Z}_2\otimes\mathbf{Z}_{\text{T}})\boldsymbol{u}_{2|1}} + \underbrace{\boldsymbol{c}\odot f_{2|1;3}(\boldsymbol{r},\boldsymbol{t})}_{((\mathbf{Z}_2\square\boldsymbol{c})\otimes\mathbf{Z}_{\text{T}})\boldsymbol{u}_{2|1;3}}}_{\text{Non-linear row and time interactions}} + \\
\underbrace{\underbrace{h_{3|1}(\boldsymbol{c},\boldsymbol{t})}_{(\mathbf{Z}_3\otimes\mathbf{Z}_{\text{T}})\boldsymbol{u}_{3|1}} + \underbrace{\boldsymbol{r}\odot f_{3|1;2}(\boldsymbol{c},\boldsymbol{t})}_{((\boldsymbol{r}\square\mathbf{Z}_3)\otimes\mathbf{Z}_{\text{T}})\boldsymbol{u}_{3|1;2}}}_{\text{Non-linear column and time interactions}} + \underbrace{\underbrace{h_{2|3|1}(\boldsymbol{r},\boldsymbol{c},\boldsymbol{t})}_{((\mathbf{Z}_2\square\mathbf{Z}_3)\otimes\mathbf{Z}_{\text{T}})\boldsymbol{u}_{2|3|1}}}_{\text{Non-linear space-time interaction}} .
\end{aligned}
\tag{5.8}
$$

Consequentely, some terms of the mixed model form (in matrix notation) of the spatio-temporal three-dimensional surface $f_{\text{ST}}(\boldsymbol{r},\boldsymbol{c},\boldsymbol{t})$ (see (3.28)) have to be removed. That is, we omitted in (3.30) the two terms, 1 and t, from $\boldsymbol{X}_{\text{ST}}$, and the Z_{T} term from $\boldsymbol{Z}_{\text{ST}}$. The new design matrices become (blocks in blue should be omitted of the following specification)

$$
\begin{aligned}
\boldsymbol{X}_{\text{ST}} &\equiv [\underbrace{\mathbf{1}_{Mn} \mid \boldsymbol{r}\otimes\mathbf{1}_n \mid \boldsymbol{c}\otimes\mathbf{1}_n \mid \boldsymbol{t}\otimes\mathbf{1}_M \mid (\boldsymbol{r}\odot\boldsymbol{c})\otimes\mathbf{1}_n \mid \boldsymbol{r}\otimes\boldsymbol{t} \mid \boldsymbol{c}\otimes\boldsymbol{t} \mid (\boldsymbol{r}\odot\boldsymbol{c})\otimes\boldsymbol{t}}_{\text{Linear effects and interactions}}], \\
\boldsymbol{Z}_{\text{ST}} &\equiv [\underbrace{\mathbf{Z}_2\otimes\mathbf{1}_n \mid (\mathbf{Z}_2\square\boldsymbol{c})\otimes\mathbf{1}_n \mid \mathbf{Z}_2\otimes\boldsymbol{t} \mid (\mathbf{Z}_2\square\boldsymbol{c})\otimes\boldsymbol{t}}_{\text{smooth row-related effects}} \mid \underbrace{\mathbf{Z}_3\otimes\mathbf{1}_n \mid (\boldsymbol{r}\square\mathbf{Z}_3)\otimes\mathbf{1}_n \mid \mathbf{Z}_3\otimes\boldsymbol{t} \mid (\boldsymbol{r}\square\mathbf{Z}_3)\otimes\boldsymbol{t}}_{\text{smooth column-related effects}} \mid \\
&\quad \underbrace{\mathbf{Z}_{\text{T}}\otimes\mathbf{1}_M \mid \boldsymbol{r}\otimes\mathbf{Z}_{\text{T}} \mid \boldsymbol{c}\otimes\mathbf{Z}_{\text{T}} \mid (\boldsymbol{r}\odot\boldsymbol{c})\otimes\mathbf{Z}_{\text{T}}}_{\text{smooth time-related effects}} \mid \underbrace{(\mathbf{Z}_2\square\mathbf{Z}_3)\otimes\mathbf{1}_M \mid (\mathbf{Z}_2\square\mathbf{Z}_3)\otimes\boldsymbol{t}}_{\text{smooth row and column interactions}} \mid \\
&\quad \underbrace{\mathbf{Z}_2\otimes\mathbf{Z}_{\text{T}} \mid (\mathbf{Z}_2\square\boldsymbol{c})\otimes\mathbf{Z}_{\text{T}}}_{\text{smooth row and time interactions}} \mid \underbrace{\mathbf{Z}_3\otimes\mathbf{Z}_{\text{T}} \mid (\boldsymbol{r}\square\mathbf{Z}_3)\otimes\mathbf{Z}_{\text{T}}}_{\text{smooth column and time interactions}} \mid \underbrace{(\mathbf{Z}_2\square\mathbf{Z}_3)\otimes\mathbf{Z}_{\text{T}}}_{\text{smooth space-time interactions}}].
\end{aligned}
\tag{5.9}
$$

In the same way, β_0 and β_3 are excluded from $\boldsymbol{\beta}_{\text{ST}}$, and $\boldsymbol{u}_1$ is removed from $\boldsymbol{u}_{\text{ST}}$ in (3.31). Thus, by excluding

the terms in blue, the vectors of coefficients are

$$\begin{aligned}\boldsymbol{\beta}_{\text{ST}} &= (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7)^T\,, \\ \boldsymbol{u}_{\text{ST}} &= \Bigg(\underbrace{\boldsymbol{u}_2^T, \boldsymbol{u}_{2;3}^T, \boldsymbol{u}_{2;1}^T, \boldsymbol{u}_{2;31}^T}_{\boldsymbol{u}_r}, \underbrace{\boldsymbol{u}_3^T, \boldsymbol{u}_{3;2}^T, \boldsymbol{u}_{3;1}^T, \boldsymbol{u}_{3;21}^T}_{\boldsymbol{u}_c}, \underbrace{\boldsymbol{u}_1^T, \boldsymbol{u}_{1;2}^T, \boldsymbol{u}_{1;3}^T, \boldsymbol{u}_{1;2,3}^T}_{\boldsymbol{u}_t} \\ &\quad \underbrace{\boldsymbol{u}_{2|3}^T, \boldsymbol{u}_{2|3;1}^T}_{\boldsymbol{u}_{r|c}}, \underbrace{\boldsymbol{u}_{2|1}^T, \boldsymbol{u}_{2|1;3}^T}_{\boldsymbol{u}_{r|t}}, \underbrace{\boldsymbol{u}_{3|1}^T, \boldsymbol{u}_{3|1,2}^T}_{\boldsymbol{u}_{c|t}}, \underbrace{\boldsymbol{u}_{2|3|1}^T}_{\boldsymbol{u}_{r|c|t}} \Bigg)^T\,.\end{aligned} \tag{5.10}$$

Notice that the vector of random effects, $\boldsymbol{u}_{\text{ST}}$, results in 18 (instead of 19) sets of random effects, each associated with one block in $\boldsymbol{Z}_{\text{ST}}$. As before, these sets can be further grouped into 7 larger sets, i.e., $\boldsymbol{u}_{\text{ST}} = (\boldsymbol{u}_r^T, \boldsymbol{u}_c^T, \boldsymbol{u}_t^T, \boldsymbol{u}_{r|c}^T, \boldsymbol{u}_{r|t}^T, \boldsymbol{u}_{c|t}^T, \boldsymbol{u}_{r|c|t}^T)$. Then, the precision matrix (i.e., the inverse of variance-covariance matrix) associated with $\boldsymbol{u}_{\text{ST}}$, is a block diagonal matrix, with 7 blocks, each related to each set of random effects

$$\begin{aligned}\boldsymbol{G}_{\text{ST}}^{-1} =& \text{blockdiag}\Bigg(\frac{1}{\sigma^2_{\text{ST},2}} \boldsymbol{\Lambda}_{2+} \otimes \boldsymbol{I}_2 \otimes \boldsymbol{I}_2, \frac{1}{\sigma^2_{\text{ST},3}} \boldsymbol{I}_2 \otimes \boldsymbol{\Lambda}_{3+} \otimes \boldsymbol{I}_2, \frac{1}{\sigma^2_{\text{ST},1}} \boldsymbol{I}_3 \otimes \boldsymbol{\Lambda}_{1+}, \\ & \frac{1}{\sigma^2_{\text{ST},2}} \boldsymbol{\Lambda}_{2+} \otimes \boldsymbol{I}_{b_3-2} \otimes \boldsymbol{I}_2 + \frac{1}{\sigma^2_{\text{ST},3}} \boldsymbol{I}_{b_2-2} \otimes \boldsymbol{\Lambda}_{3+} \otimes \boldsymbol{I}_2, \\ & \frac{1}{\sigma^2_{\text{ST},2}} \boldsymbol{\Lambda}_{2+} \otimes \boldsymbol{I}_2 \otimes \boldsymbol{I}_{b_1-2} + \frac{1}{\sigma^2_{\text{ST},1}} \boldsymbol{I}_{b_2-2} \otimes \boldsymbol{I}_2 \otimes \boldsymbol{\Lambda}_{1+}, \\ & \frac{1}{\sigma^2_{\text{ST},3}} \boldsymbol{I}_2 \otimes \boldsymbol{\Lambda}_{3+} \otimes \boldsymbol{I}_{b_1-2} + \frac{1}{\sigma^2_{\text{ST},1}} \boldsymbol{I}_2 \otimes \boldsymbol{I}_{b_3-2} \otimes \boldsymbol{\Lambda}_{1+}, \\ & \frac{1}{\sigma^2_{\text{ST},2}} \boldsymbol{\Lambda}_{2+} \otimes \boldsymbol{I}_{b_3-2} \otimes \boldsymbol{I}_{b_1-2} + \frac{1}{\sigma^2_{\text{ST},3}} \boldsymbol{I}_{b_2-2} \otimes \boldsymbol{\Lambda}_{3+} \otimes \boldsymbol{I}_{b_1-2} + \frac{1}{\sigma^2_{\text{ST},1}} \boldsymbol{I}_{b_2-2} \otimes \boldsymbol{I}_{b_3-2} \otimes \boldsymbol{\Lambda}_{1+} \Bigg) \\ =& \sum_{j \in \{2,3,1\}} \sigma^{-2}_{\text{ST},j} \tilde{\boldsymbol{\Lambda}}_{j+},\end{aligned} \tag{5.11}$$

where the specific form of $\boldsymbol{\Lambda}_{j+}$ and $\tilde{\boldsymbol{\Lambda}}_{j+}$ ($j = 1, 2, 3$) are given in Section 3.2.3 (once again it is obtained from the connection between P-splines and linear mixed models). We note that the difference between the $\boldsymbol{G}_{\text{ST}}^{-1}$ in (3.32) and (5.11) is the (blue) block associated with the one-dimensional smooth (non-linear) effects along time ($\boldsymbol{u}_t$); in this case, we have removed the non-linear main effect along time ($\boldsymbol{u}_1$). As can be observed, $\boldsymbol{G}_{\text{ST}}^{-1}$ depends on three variance parameters, $\sigma^2_{\text{ST},1}$, $\sigma^2_{\text{ST},2}$, and $\sigma^2_{\text{ST},3}$; they are responsible for controlling the smoothness along the rows, columns and time, respectively. Consequently, $\tilde{\boldsymbol{\Lambda}}_{1+}$ in (3.33) becomes

$$\tilde{\boldsymbol{\Lambda}}_{1+} = \text{blockdiag}\left(\boldsymbol{0}, \boldsymbol{0}, \boldsymbol{I}_2 \otimes \boldsymbol{I}_2 \otimes \boldsymbol{\Lambda}_{1+}, \boldsymbol{0}, \boldsymbol{I}_{b_2-2} \otimes \boldsymbol{I}_2 \otimes \boldsymbol{\Lambda}_{1+}, \boldsymbol{I}_2 \otimes \boldsymbol{I}_{b_3-2} \otimes \boldsymbol{\Lambda}_{1+}, \boldsymbol{I}_3 \otimes \boldsymbol{\Lambda}_{1+}\right).$$

With all ingredients introduced before (i.e., the specifications in equations (5.7) for the longitudinal one-dimensional functions, and in equation (5.9) for the spatio-temporal function), in matrix notation model

(5.4) is expressed, for all M plants, as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\varepsilon},\ \boldsymbol{u} \sim N(\boldsymbol{0}, \boldsymbol{G}),\ \boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I}), \tag{5.12}$$

where

$$\begin{aligned}
\boldsymbol{X} &= [\boldsymbol{Q}_{\text{pop}} \otimes \boldsymbol{X}_{\text{pop}} \mid \boldsymbol{X}_{\text{ST}}],\\
\boldsymbol{Z} &= [\boldsymbol{Q}_{\text{pop}} \otimes \boldsymbol{Z}_{\text{pop}} \mid \boldsymbol{Q}_{\text{gen}} \otimes \boldsymbol{X}_{\text{gen}} \mid \boldsymbol{Q}_{\text{gen}} \otimes \boldsymbol{Z}_{\text{gen}} \mid \boldsymbol{I}_M \otimes \boldsymbol{X}_{\text{plant}} \mid \boldsymbol{I}_M \otimes \boldsymbol{Z}_{\text{plant}} \mid\\
&\quad \boldsymbol{Q}_{\text{row}} \otimes \boldsymbol{X}_{\text{row}} \mid \boldsymbol{Q}_{\text{row}} \otimes \boldsymbol{Z}_{\text{row}} \mid \boldsymbol{Q}_{\text{col}} \otimes \boldsymbol{X}_{\text{col}} \mid \boldsymbol{Q}_{\text{col}} \otimes \boldsymbol{Z}_{\text{col}} \mid \boldsymbol{Z}_{\text{ST}}],
\end{aligned}$$

where the contrast matrices $\boldsymbol{Q}_{\text{pop}}$, $\boldsymbol{Q}_{\text{gen}}$, $\boldsymbol{Q}_{\text{row}}$ and $\boldsymbol{Q}_{\text{col}}$ were defined before in (5.4), and

$$\begin{aligned}
\boldsymbol{\beta} &= \left(\boldsymbol{\beta}_{\text{pop}}^T, \boldsymbol{\beta}_{\text{ST}}^T\right)^T\\
\boldsymbol{u} &= (\boldsymbol{u}_{\text{nlin, pop}}^T, \boldsymbol{u}_{\text{lin, gen}}^T, \boldsymbol{u}_{\text{nlin, gen}}^T, \boldsymbol{u}_{\text{lin, plant}}^T, \boldsymbol{u}_{\text{nlin, plant}}^T, \boldsymbol{u}_{\text{lin, row}}^T, \boldsymbol{u}_{\text{nlin, row}}^T, \boldsymbol{u}_{\text{lin, col}}^T, \boldsymbol{u}_{\text{nlin, col}}^T, \boldsymbol{u}_{\text{ST}}^T)^T,
\end{aligned}$$

where

$$\begin{aligned}
\boldsymbol{\beta}_{\text{pop}} &= \left(\boldsymbol{\beta}_{\text{lin},1}^{\text{pop};T}, \ldots, \boldsymbol{\beta}_{\text{lin},K}^{\text{pop};T}\right)^T, \quad \boldsymbol{u}_{\text{nlin, pop}} = \left(\boldsymbol{u}_{\text{nlin},1}^{\text{pop};T}, \ldots, \boldsymbol{u}_{\text{nlin},K}^{\text{pop};T}\right)^T,\\
\boldsymbol{u}_{\text{lin, gen}} &= \left(\boldsymbol{u}_{\text{lin},1}^{\text{gen};T}, \ldots, \boldsymbol{u}_{\text{lin},L}^{\text{gen};T}\right)^T, \quad \boldsymbol{u}_{\text{nlin, gen}} = \left(\boldsymbol{u}_{\text{nlin},1}^{\text{gen};T}, \ldots, \boldsymbol{u}_{\text{nlin},L}^{\text{gen};T}\right)^T,\\
\boldsymbol{u}_{\text{lin, plant}} &= \left(\boldsymbol{u}_{\text{lin},1}^{\text{plant};T}, \ldots, \boldsymbol{u}_{\text{lin},M}^{\text{plant};T}\right)^T, \boldsymbol{u}_{\text{nlin, plant}} = \left(\boldsymbol{u}_{\text{nlin},1}^{\text{plant};T}, \ldots, \boldsymbol{u}_{\text{nlin},M}^{\text{plant};T}\right)^T,\\
\boldsymbol{u}_{\text{lin, row}} &= \left(\boldsymbol{u}_{\text{lin},1}^{\text{row};T}, \ldots, \boldsymbol{u}_{\text{lin},R}^{\text{row};T}\right)^T, \quad \boldsymbol{u}_{\text{nlin, row}} = \left(\boldsymbol{u}_{\text{nlin},1}^{\text{row};T}, \ldots, \boldsymbol{u}_{\text{nlin},R}^{\text{row};T}\right)^T,\\
\boldsymbol{u}_{\text{lin, col}} &= \left(\boldsymbol{u}_{\text{lin},1}^{\text{col};T}, \ldots, \boldsymbol{u}_{\text{lin},C}^{\text{col};T}\right)^T, \quad \boldsymbol{u}_{\text{nlin, col}} = \left(\boldsymbol{u}_{\text{nlin},1}^{\text{col};T}, \ldots, \boldsymbol{u}_{\text{nlin},C}^{\text{col};T}\right)^T.
\end{aligned}$$

Finally, the variance-covariance matrix $\boldsymbol{G}$ is a block diagonal matrix given by

$$\boldsymbol{G} = \text{blockdiag}(\boldsymbol{G}_{\text{pop}}, \boldsymbol{G}_{\text{geno}}, \boldsymbol{G}_{\text{plant}}, \boldsymbol{G}_{\text{row}}, \boldsymbol{G}_{\text{col}}, \boldsymbol{G}_{\text{ST}}), \tag{5.13}$$

where

$$\begin{aligned}
\boldsymbol{G}_{\text{pop}} &= \text{blockdiag}\left(\sigma_{\text{pop},1}^2\boldsymbol{\Sigma}_{\text{pop+}}, \ldots, \sigma_{\text{pop},K}^2\boldsymbol{\Sigma}_{\text{pop+}}\right),\\
\boldsymbol{G}_{\text{geno}} &= \text{blockdiag}\left(\boldsymbol{I}_L \otimes \boldsymbol{\Sigma}_{\text{gen}}, \boldsymbol{I}_L \otimes \sigma_{\text{gen}}^2\boldsymbol{\Sigma}_{\text{gen+}}\right),\\
\boldsymbol{G}_{\text{plant}} &= \text{blockdiag}\left(\boldsymbol{I}_M \otimes \boldsymbol{\Sigma}_{\text{plant}}, \boldsymbol{I}_M \otimes \sigma_{\text{plant}}^2\boldsymbol{\Sigma}_{\text{plant+}}\right),\\
\boldsymbol{G}_{\text{row}} &= \text{blockdiag}\left(\boldsymbol{I}_R \otimes \boldsymbol{\Sigma}_{\text{row}}, \boldsymbol{I}_R \otimes \sigma_{\text{row}}^2\boldsymbol{\Sigma}_{\text{row+}}\right),\\
\boldsymbol{G}_{\text{col}} &= \text{blockdiag}\left(\boldsymbol{I}_C \otimes \boldsymbol{\Sigma}_{\text{col}}, \boldsymbol{I}_C \otimes \sigma_{\text{col}}^2\boldsymbol{\Sigma}_{\text{col+}}\right).
\end{aligned}$$

We note that the variance-covariance (5.13) has a non-standard form, with the last block in $\boldsymbol{G}_{\text{ST}}$ depending on three variance parameters in a non linear way. More precisely, in our case what is linear in the (inverse of the) variance parameters is the precision matrix, $\boldsymbol{G}^{-1}$ and not the variance-covariance matrix, $\boldsymbol{G}$, i.e.,

$\boldsymbol{G}^{-1} = \sum_j \sigma_j^{-2} \tilde{\boldsymbol{\Lambda}}_j$ where

$$\begin{aligned}
\tilde{\boldsymbol{\Lambda}}_1 &= \text{blockdiag}(\boldsymbol{\Lambda}_{\text{pop+}}, \ldots, \boldsymbol{0}, \boldsymbol{I}_L \otimes \text{blockdiag}(0,0,\boldsymbol{0}), \boldsymbol{I}_M \otimes \text{blockdiag}(0,0,\boldsymbol{0}),\\
&\qquad \boldsymbol{I}_R \otimes \text{blockdiag}(0,0,\boldsymbol{0}), \boldsymbol{I}_C \otimes \text{blockdiag}(0,0,\boldsymbol{0}), \boldsymbol{0}),\\
&\vdots\\
\tilde{\boldsymbol{\Lambda}}_K &= \text{blockdiag}(\boldsymbol{0}, \ldots, \boldsymbol{\Lambda}_{\text{pop+}}, \boldsymbol{I}_L \otimes \text{blockdiag}(0,0,\boldsymbol{0}), \boldsymbol{I}_M \otimes \text{blockdiag}(0,0,\boldsymbol{0}),\\
&\qquad \boldsymbol{I}_R \otimes \text{blockdiag}(0,0,\boldsymbol{0}), \boldsymbol{I}_C \otimes \text{blockdiag}(0,0,\boldsymbol{0}), \boldsymbol{0}),
\end{aligned}$$

are block diagonal matrices with zeroes matrices of proper dimension each, associated with the effects at the population level,

$$\begin{aligned}
\tilde{\boldsymbol{\Lambda}}_{K+1} &= \text{blockdiag}(\boldsymbol{0}, \ldots, \boldsymbol{0}, \boldsymbol{I}_L \otimes \text{blockdiag}(1,0,\boldsymbol{0}), \boldsymbol{I}_M \otimes \text{blockdiag}(0,0,\boldsymbol{0}),\\
&\qquad \boldsymbol{I}_R \otimes \text{blockdiag}(0,0,\boldsymbol{0}), \boldsymbol{I}_C \otimes \text{blockdiag}(0,0,\boldsymbol{0}), \boldsymbol{0}),\\
\tilde{\boldsymbol{\Lambda}}_{K+2} &= \text{blockdiag}(\boldsymbol{0}, \ldots, \boldsymbol{0}, \boldsymbol{I}_L \otimes \text{blockdiag}(0,1,\boldsymbol{0}), \boldsymbol{I}_M \otimes \text{blockdiag}(0,0,\boldsymbol{0}),\\
&\qquad \boldsymbol{I}_R \otimes \text{blockdiag}(0,0,\boldsymbol{0}), \boldsymbol{I}_C \otimes \text{blockdiag}(0,0,\boldsymbol{0}), \boldsymbol{0}),\\
\tilde{\boldsymbol{\Lambda}}_{K+3} &= \text{blockdiag}(\boldsymbol{0}, \ldots, \boldsymbol{0}, \boldsymbol{I}_L \otimes \text{blockdiag}(0,0,\boldsymbol{\Lambda}_{\text{gen+}}), \boldsymbol{I}_M \otimes \text{blockdiag}(0,0,\boldsymbol{0}),\\
&\qquad \boldsymbol{I}_R \otimes \text{blockdiag}(0,0,\boldsymbol{0}), \boldsymbol{I}_C \otimes \text{blockdiag}(0,0,\boldsymbol{0}), \boldsymbol{0}),
\end{aligned}$$

are block diagonal matrices with zeroes matrices of proper dimension each, associated with the effects at the genotype level,

$$\begin{aligned}
\tilde{\boldsymbol{\Lambda}}_{K+4} &= \text{blockdiag}(\boldsymbol{0}, \ldots, \boldsymbol{0}, \boldsymbol{I}_L \otimes \text{blockdiag}(0,0,\boldsymbol{0}), \boldsymbol{I}_M \otimes \text{blockdiag}(1,0,\boldsymbol{0}),\\
&\qquad \boldsymbol{I}_R \otimes \text{blockdiag}(0,0,\boldsymbol{0}), \boldsymbol{I}_C \otimes \text{blockdiag}(0,0,\boldsymbol{0}), \boldsymbol{0}),\\
\tilde{\boldsymbol{\Lambda}}_{K+5} &= \text{blockdiag}(\boldsymbol{0}, \ldots, \boldsymbol{0}, \boldsymbol{I}_L \otimes \text{blockdiag}(0,0,\boldsymbol{0}), \boldsymbol{I}_M \otimes \text{blockdiag}(0,1,\boldsymbol{0}),\\
&\qquad \boldsymbol{I}_R \otimes \text{blockdiag}(0,0,\boldsymbol{0}), \boldsymbol{I}_C \otimes \text{blockdiag}(0,0,\boldsymbol{0}), \boldsymbol{0}),\\
\tilde{\boldsymbol{\Lambda}}_{K+6} &= \text{blockdiag}(\boldsymbol{0}, \ldots, \boldsymbol{0}, \boldsymbol{I}_L \otimes \text{blockdiag}(0,0,\boldsymbol{0}), \boldsymbol{I}_M \otimes \text{blockdiag}(0,0,\boldsymbol{\Lambda}_{\text{plant+}}),\\
&\qquad \boldsymbol{I}_R \otimes \text{blockdiag}(0,0,\boldsymbol{0}), \boldsymbol{I}_C \otimes \text{blockdiag}(0,0,\boldsymbol{0}), \boldsymbol{0}),
\end{aligned}$$

are block diagonal matrices with zeroes matrices of proper dimension each, associated with the effects at the plant level,

$$\begin{aligned}
\tilde{\boldsymbol{\Lambda}}_{K+7} &= \text{blockdiag}(\boldsymbol{0}, \ldots, \boldsymbol{0}, \boldsymbol{I}_L \otimes \text{blockdiag}(0,0,\boldsymbol{0}), \boldsymbol{I}_M \otimes \text{blockdiag}(0,0,\boldsymbol{0}),\\
&\qquad \boldsymbol{I}_R \otimes \text{blockdiag}(1,0,\boldsymbol{0}), \boldsymbol{I}_C \otimes \text{blockdiag}(0,0,\boldsymbol{0}), \boldsymbol{0}),\\
\tilde{\boldsymbol{\Lambda}}_{K+8} &= \text{blockdiag}(\boldsymbol{0}, \ldots, \boldsymbol{0}, \boldsymbol{I}_L \otimes \text{blockdiag}(0,0,\boldsymbol{0}), \boldsymbol{I}_M \otimes \text{blockdiag}(0,0,\boldsymbol{0}),\\
&\qquad \boldsymbol{I}_R \otimes \text{blockdiag}(0,1,\boldsymbol{0}), \boldsymbol{I}_C \otimes \text{blockdiag}(0,0,\boldsymbol{0}), \boldsymbol{0}),\\
\tilde{\boldsymbol{\Lambda}}_{K+9} &= \text{blockdiag}(\boldsymbol{0}, \ldots, \boldsymbol{0}, \boldsymbol{I}_L \otimes \text{blockdiag}(0,0,\boldsymbol{0}), \boldsymbol{I}_M \otimes \text{blockdiag}(0,0,\boldsymbol{0}),\\
&\qquad \boldsymbol{I}_R \otimes \text{blockdiag}(0,0,\boldsymbol{\Lambda}_{\text{row+}}), \boldsymbol{I}_C \otimes \text{blockdiag}(0,0,\boldsymbol{0}), \boldsymbol{0}),
\end{aligned}$$

are block diagonal matrices with zeroes matrices of proper dimension each, associated with the row effects,

$$\begin{aligned}
\tilde{\boldsymbol{\Lambda}}_{K+10} &= \text{blockdiag}(\mathbf{0}, \ldots, \mathbf{0}, \boldsymbol{I}_L \otimes \text{blockdiag}(0, 0, \mathbf{0}), \boldsymbol{I}_M \otimes \text{blockdiag}(0, 0, \mathbf{0}), \\
&\qquad \boldsymbol{I}_R \otimes \text{blockdiag}(0, 0, \mathbf{0}), \boldsymbol{I}_C \otimes \text{blockdiag}(1, 0, \mathbf{0}), \mathbf{0}), \\
\tilde{\boldsymbol{\Lambda}}_{K+11} &= \text{blockdiag}(\mathbf{0}, \ldots, \mathbf{0}, \boldsymbol{I}_L \otimes \text{blockdiag}(0, 0, \mathbf{0}), \boldsymbol{I}_M \otimes \text{blockdiag}(0, 0, \mathbf{0}), \\
&\qquad \boldsymbol{I}_R \otimes \text{blockdiag}(0, 0, \mathbf{0}), \boldsymbol{I}_C \otimes \text{blockdiag}(0, 1, \mathbf{0}), \mathbf{0}), \\
\tilde{\boldsymbol{\Lambda}}_{K+12} &= \text{blockdiag}(\mathbf{0}, \ldots, \mathbf{0}, \boldsymbol{I}_L \otimes \text{blockdiag}(0, 0, \mathbf{0}), \boldsymbol{I}_M \otimes \text{blockdiag}(0, 0, \mathbf{0}), \\
&\qquad \boldsymbol{I}_R \otimes \text{blockdiag}(0, 0, \mathbf{0}), \boldsymbol{I}_C \otimes \text{blockdiag}(0, 0, \boldsymbol{\Lambda}_{\text{col}+}), \mathbf{0}),
\end{aligned}$$

are block diagonal matrices with zeroes matrices of proper dimension each, associated with the column effect,

$$\begin{aligned}
\tilde{\boldsymbol{\Lambda}}_{K+13} &= \text{blockdiag}(\mathbf{0}, \ldots, \mathbf{0}, \boldsymbol{I}_L \otimes \text{blockdiag}(0, 0, \mathbf{0}), \boldsymbol{I}_M \otimes \text{blockdiag}(0, 0, \mathbf{0}), \\
&\qquad \boldsymbol{I}_R \otimes \text{blockdiag}(0, 0, \mathbf{0}), \boldsymbol{I}_C \otimes \text{blockdiag}(0, 0, \mathbf{0}), \tilde{\boldsymbol{\Lambda}}_{2+}), \\
\tilde{\boldsymbol{\Lambda}}_{K+14} &= \text{blockdiag}(\mathbf{0}, \ldots, \mathbf{0}, \boldsymbol{I}_L \otimes \text{blockdiag}(0, 0, \mathbf{0}), \boldsymbol{I}_M \otimes \text{blockdiag}(0, 0, \mathbf{0}), \\
&\qquad \boldsymbol{I}_R \otimes \text{blockdiag}(0, 0, \mathbf{0}), \boldsymbol{I}_C \otimes \text{blockdiag}(0, 0, \mathbf{0}), \tilde{\boldsymbol{\Lambda}}_{3+}), \\
\tilde{\boldsymbol{\Lambda}}_{K+15} &= \text{blockdiag}(\mathbf{0}, \ldots, \mathbf{0}, \boldsymbol{I}_L \otimes \text{blockdiag}(0, 0, \mathbf{0}), \boldsymbol{I}_M \otimes \text{blockdiag}(0, 0, \mathbf{0}), \\
&\qquad \boldsymbol{I}_R \otimes \text{blockdiag}(0, 0, \mathbf{0}), \boldsymbol{I}_C \otimes \text{blockdiag}(0, 0, \mathbf{0}), \tilde{\boldsymbol{\Lambda}}_{1+}),
\end{aligned}$$

are block diagonal matrices with zeroes matrices of proper dimension each, associated with the spatio-temporal effects, where $\tilde{\boldsymbol{\Lambda}}_{1+}$ was redefined in (5.2), and $\tilde{\boldsymbol{\Lambda}}_{2+}$ and $\tilde{\boldsymbol{\Lambda}}_{3+}$ were defined in (3.33).

Finally, to be aware of the model complexity, we note that under this configuration, the mixed model formulation of our one-stage approach has a total of

- $K + L + M + R + C$ one-dimensional smooth functions (populations + genotypes + plants + rows + columns), plus the spatio-temporal smooth function (in the row, column, and time directions),
- $K \times b_{\text{pop}} + L \times b_{\text{gen}} + M \times b_{\text{plant}} + R \times b_{\text{row}} + C \times b_{\text{col}}$ regression coefficients (both, fixed and random) for the one-dimensional smooth functions, and $(b_2 b_3 - 1) b_1$ regression coefficients (both, fixed and random, and by considering second-order difference penalties and cubic B-spline basis) associated with the spatio-temporal smooth function, and
- $K + 16$ variance parameters:
  - K variance parameters at population level, $\sigma^2_1 = \sigma^2_{\text{pop},1}, \ldots, \sigma^2_K = \sigma^2_{\text{pop},K}$,
  - 3 variance parameters at genotype level, $\sigma^2_{K+1} = \sigma^2_{\text{gen},0}$, $\sigma^2_{K+2} = \sigma^2_{\text{gen},1}$, and $\sigma^2_{K+3} = \sigma^2_{\text{gen}}$,
  - 3 variance parameters at plant level, $\sigma^2_{K+4} = \sigma^2_{\text{plant},0}$, $\sigma^2_{K+5} = \sigma^2_{\text{plant},1}$, and $\sigma^2_{K+6} = \sigma^2_{\text{plant}}$,
  - 3 variance parameters for the row effects, $\sigma^2_{K+7} = \sigma^2_{\text{row},0}$, $\sigma^2_{K+8} = \sigma^2_{\text{row},1}$, and $\sigma^2_{K+9} = \sigma^2_{\text{row}}$,

- 3 variance parameters for the column effects, $\sigma^2_{K+10} = \sigma^2_{\text{col},0}$, $\sigma^2_{K+11} = \sigma^2_{\text{col},1}$, and $\sigma^2_{K+12} = \sigma^2_{\text{col}}$,
- 3 variance parameters for the spatio-temporal smooth function (in the row, column and time directions), $\sigma^2_{K+13} = \sigma^2_{\text{ST},1}$, $\sigma^2_{K+14} = \sigma^2_{\text{ST},2}$, and $\sigma^2_{K+15} = \sigma^2_{\text{ST},3}$, and
- the residual variance, σ^2.

Besides, as we have proposed a P-spline-based approach, the computation time increases with the number of plants and basis dimensions, resulting in a scalability problem. Based on our experience, we recommend using the same number of B-spline basis functions for the three levels of the hierarchy (b_{pop}, b_{gen} and b_{plant}), as well as for the row and column random effects (b_{row} and b_{col}). Regarding the number of B-spline basis functions used for the three-dimensional surface (b_1, b_2 and b_3), we suggest keeping them relatively small to enable the solution to run on standard computers.

5.3 Computational aspects

Implementation of our one-stage approach is even more challenging than the previous two-stage approach proposed. Here, all the complexity is addressed with a single model increasing the computational burden. Thus, to make our one-stage approach computationally affordable, we combine different specialised methods that take advantage of the sparse model matrices structure (sparse matrix algebra in the `R`-packages `spam` and `Matrix`, and the Linear Mixed Model Solver `LMMsolver`), the array data structure (GLAM), and the non-standard form of the variance-covariance matrix (SOP). Although the SOP algorithm can be used to estimate any of the models proposed in this thesis, it is especially relevant for our one-stage approach (the variance-covariance matrix $\boldsymbol{G}$ in (5.13) does really have a non-standard form). We follow the same ideas as in Section 4.2.5 to estimate our one-stage approach. That is, we integrate the computational tools indicated before to the SOP algorithm in Section 3.3, where BLUEs and BLUPs are obtained by the solution of Henderson's mixed model equations, and variance parameters are estimated using REML. As for the two-stage approach, we implemented our own code in the `R` language to keep the computational effort manageable, which will be publicly available through the `statgenHTP` `R`-package (Millet et al., 2022), and that it is now accessible on https://gitlab.bcamath.org/dperez/htp_one_stage_approach. In Chapter 8, we describe details on the usage of the developed code.

5.4 Derivatives, standard errors and pointwise confidence intervals

As result of the one-stage approach, we can obtain estimated curves (trajectories and deviations) at the three levels of the hierarchy (populations, genotypes and plants; in the same way that are obtained for our two-stage approach) plus estimated curves (deviations) for rows and columns, and an estimated three-

dimensional surface in the row, column and time directions. However, in the application examples in Chapter 7, we are particularly interested in the estimated curves for the hierarchical structure. For these hierarchical estimated curves, we can further estimate first-order derivatives and confidence intervals as addressed in Sections 3.1.1.1 (for derivatives estimation), 3.4 (for confidence intervals estimation), and 4.2.6 (for inference procedures used in the two-stage approach). Once again, the computational challenge (and limitation) at this point is to obtain the inverse of the variance-covariance matrix, $\boldsymbol{C}^{-1}$, to calculate standard errors. We note that the size of this matrix is much larger than the one obtained in the second stage of the two-stage approach. Consequentely, confidence intervals estimation is very time-consuming and memory-intensive. We believe one solution is to obtain standard errors without calculating $\boldsymbol{C}^{-1}$ in a similar way that effective dimensions are estimated using the `LMMsolver`.

5.5 Time-independent features extraction

The aim here is to extract time-independent characteristics (intermediate traits) from the estimated curves (trajectories, deviations and first-order derivatives) at the three hierarchy levels (populations, genotypes and plants) for further genotype analysis. As is proposed for the two-stage approach in Section 4.2.7 and depicted in Figure 4.2, some examples are the maximum and minimum trait (from trajectories), the maximum and average growth rate (from the first-order derivative of the trajectories), the onset and end of senescence/growth (from second-order derivatives of the trajectories), and the area under the curve (AUC, from the deviation curves).

Chapter 6

Simulation study

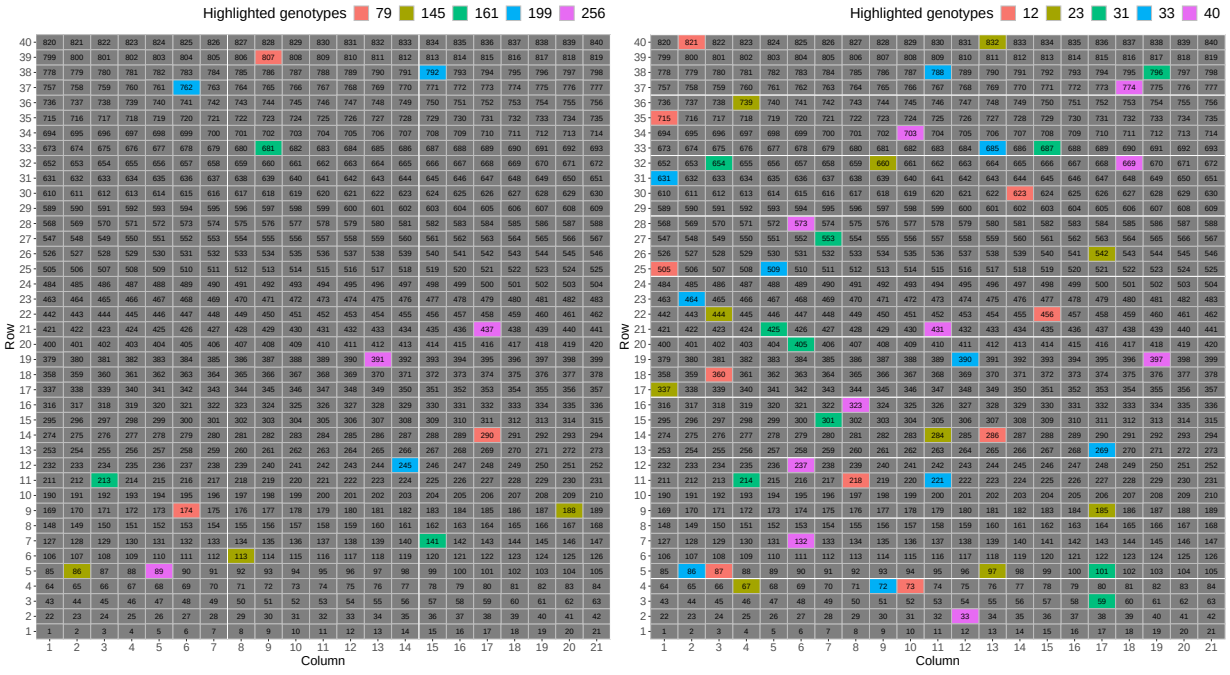
When applying the one-stage (Chapter 5) and two-stage (Chapter 4) approaches to real data sets, two problems arise: (i) the phenotypic trait is only measured at the plant level, with no available measurements at the population and genotype levels; as a result, it is not feasible to evaluate models' performance in estimating curves at these higher levels using real data, and (ii) the approaches seem to be sensitive to the dimension of the B-spline bases used at each level of the hierarchy. In this chapter, we present a data generating model and a simulation experiment to study the above-mentioned problems. The data generating model aims to mimic the spatio-temporal three-nested hierarchical data structure that motivated this thesis while keeping it independent from the statistical models used for their analysis. The generating model decomposes the spatio-temporal variation of the phenotype of interest in three components (for simplicity, we consider one population): within genotypes and plant variation, and spatio-temporal correlated noise. Data are simulated by considering the between genotype and plant (deviations) variability, and the number of replicates per genotype effects. The results of the simulations assess the performance of the one- and two-stage approaches and five different configurations for the dimensions of the B-spline bases for the hierarchical components. Ultimately, this simulation study aims to illustrate the strengths and limitations of the two proposed approaches.

6.1 Data generating model

Following the same notation introduced in the previous chapters, we simulate HTP data assuming the following spatio-temporal three-level nested hierarchical model

$$y_i(t) = f_{p(i)}(t) + f_{g(i)}(t) + f_i(t) + \varepsilon(r(i), c(i), t), \tag{6.1}$$

where $1 \leq g \leq L$, $1 \leq i \leq M$, $1 \leq r \leq R$ $1 \leq c \leq C$, and $t \in \{t_1, \ldots, t_n\}$. For simplicity, here we consider only one population ($K = 1$), and $M = R \times C$ plants (i.e., the number of plants corresponds to the number of spatial locations, with a 1:1 correspondence). In what follows we assume that the plants are located on the $R \times C$ grid such that the first plant ($i = 1$) is in the first row and column position (i.e., $r(1) = 1$ and $c(1) = 1$), and the last plant ($i = M$) is in the last row and column position of the grid (i.e., $r(M) = R$ and $c(M) = C$), and plants are ordered by rows. Figure 6.1 depicts, for two simulated datasets (each with different number of replicates by genotypes, i.e., $m_g = 3$ and 10) the grid with the randomisation used (complete block design).



(a) $m_g = 3$ (b) $m_g = 10$

Figure 6.1: Illustrative visualisation of the grid with the randomisation used for two simulated datasets (as illustration) with (**a**) $m_g = 3$, and (**b**) $m_g = 10$ replicates per genotype. The size of the grid is $R \times C = 40 \times 21$, for a total of $M = 840$ plants. Each cell represents a plant ($i = 1, \ldots, M$), which is identified by its row and column position, i.e., the ith plant has coordinates $(r(i), c(i))$ (e.g., plant $i = 457$ has position $(r(457), c(457)) = (22, 16)$). Colours depict replicates in five selected genotypes (as illustration).

We note that the data generating model, equation (6.1), is independent from the statistical models,

(SpATS model (4.1) and the psHDM (4.6) for the two-stage approach, and the spatio-temporal psHDM (5.2) for the one-stage approach), used for the analysis. The difference between the three models lies in how the spatio-temporal (stochastic) component is incorporated. In the data generating model (6.1), we have spatio-temporal correlated noise, $\varepsilon(r, c, t)$; in the spatio-temporal psHDM (5.2), we have a spatio-temporal three-dimensional surface, $f_{\text{ST}}(r, c, t)$, and the pure random noise, ε_i; while in the two-stage approach the spatial and (environmental) temporal components are modeled independently in the first stage, i.e., the spatial trend is modelled through $h_{\text{S}}(r, c)$ in the SpATS model (4.1), separately for each time point. Besides, the data generating model (6.1) does not consider the row and column random effects. However, what is common to the three models, (4.6), (5.2), and (6.1), is the three-level hierarchical structure.

In particular, the data are generated from the population to the plant level in the following five steps (we provide Figure 6.2 to graphically summarise the procedure and the kind of curves that are obtained at each step)

Step 1. Generate one population trajectory, $\boldsymbol{f}_p = (f_p(t_1), \ldots, f_p(t_n))^T$, from the growth logistic curve model. Following the notation in Z. Li and Sillanpää (2015), we consider

$$\boldsymbol{f}_p = \frac{a}{1 + e^{c(b-\boldsymbol{t})}},$$

where a is the asymptote, b is the inflection point, and c is the growth rate. Additionally, it can be shown that the first-order derivative of this function with respect to time, t, is

$$\boldsymbol{f}'_p = \frac{ace^{c(b-\boldsymbol{t})}}{(1 + e^{c(b-\boldsymbol{t})})^2}.$$

Step 2. Generate L genotype-specific deviations, $\boldsymbol{f}_g = (f_g(t_1), \ldots, f_g(t_n))^T \overset{\text{iid}}{\sim} N(\boldsymbol{0}, \boldsymbol{\Sigma}_{\text{gen}})$ $(g = 1, \ldots, L)$. For the $n \times n$ variance-covariance matrix $\boldsymbol{\Sigma}_{\text{gen}}$, we consider a first-order autoregressive structure with heterogeneous variance (ARH(1), with a slight modification to the structure presented by Wolfinger, 1996, to account for variance increasing with time structure), that is

$$\left(\boldsymbol{\Sigma}_{\text{gen}}\right)_{jk} = \frac{1}{1 - \rho^2} s_{jk}^{\text{gen}} \rho^{d_{jk}},$$

where ρ is the autocorrelation parameter, d_{jk} is the euclidean distance between time points t_j and t_k (i.e., $d_{jk} = |t_k - t_j|$)), and $s_{jk}^{\text{gen}} = \sigma_{\text{gen}}^4 h(t_j) h(t_k)$ are the elements of the heterogeneous variance-covariance matrix. Here, σ_{gen}^2 is the between genotype (deviation) variability, and $h(\cdot)$ is a function that is quadratic in t. The genotype trajectories can be obtained as the sum of the population trajectory and the genotype deviations, i.e., $\boldsymbol{f}_p + \boldsymbol{f}_{g(p)}$, as depicted in Figure 6.2.

Step 3. Generate M $(= R \times C)$ plant-specific deviations, $\boldsymbol{f}_i = (f_i(t_1), \ldots, f_i(t_n))^T \overset{\text{iid}}{\sim} N(\boldsymbol{0}, \boldsymbol{\Sigma}_{\text{plant}})$ $(i = 1, \ldots, M)$. We follow the same ideas used in **Step 2** to obtain the $n \times n$ variance-covariance matrix $\boldsymbol{\Sigma}_{\text{plant}}$. We use σ_{plant}^2 for the between plants (deviation) variability, and $s_{jk}^{\text{plant}} = \sigma_{\text{plant}}^4 h(t_j) h(t_k)$ for the elements of the heterogeneous variance-covariance matrix.

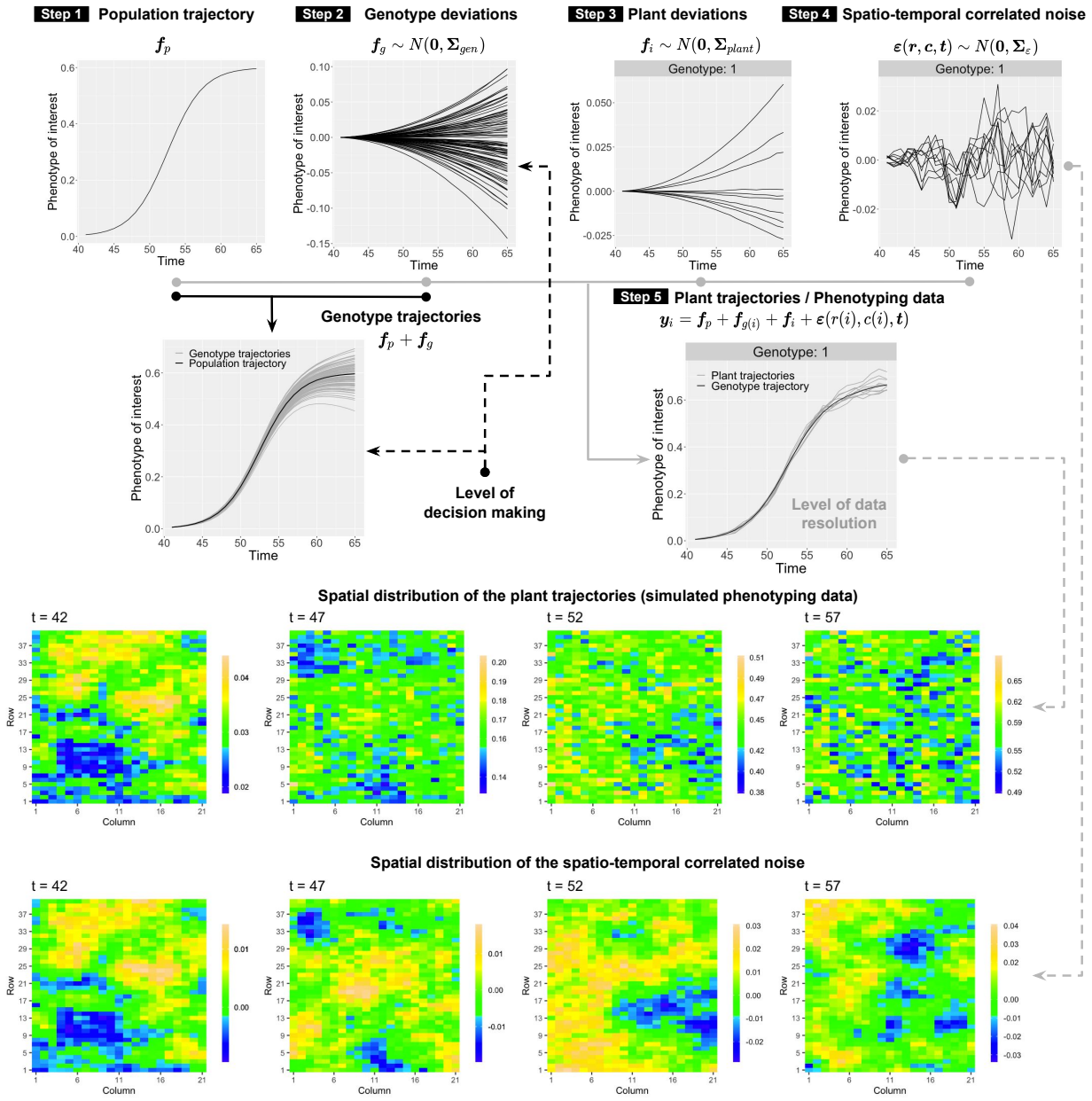


Figure 6.2: For the simulation study: Data generating strategy. Reference values to simulate the *Phenotype of interest* are based on the leaf area (m^2 plant^{-1}) data from the PhenoArch platform as described in the simulation settings on Table 6.1.

Step 4. Generate $M = R \times C$ spatio-temporal correlated noise curves, $\boldsymbol{\varepsilon}(\boldsymbol{r}, \boldsymbol{c}, \boldsymbol{t}) = (\varepsilon(r(1), c(1), \boldsymbol{t}), \varepsilon(r(2), c(2), \boldsymbol{t}), \ldots, \varepsilon(r(M), c(M), \boldsymbol{t}))^T \sim N(\mathbf{0}, \boldsymbol{\Sigma}_\varepsilon)$, where $\varepsilon(r(i), c(i), \boldsymbol{t}) = (\varepsilon(r(i), c(i), t_1), \ldots, \varepsilon(r(i), c(i), t_n))$ is the noise curve for the ith plant. For the $(RCn) \times (RCn)$ variance-covariance matrix $\boldsymbol{\Sigma}_\varepsilon$, we use a space-time separable covariance model $\boldsymbol{\Sigma}_\varepsilon = \boldsymbol{\Sigma}_\mathrm{S} \otimes \boldsymbol{\Sigma}_\mathrm{T}$, where $\boldsymbol{\Sigma}_\mathrm{S}$ is a $(RC) \times (RC)$ isotropic and homogeneous spatial variance-covariance matrix (Matérn, Guttorp & Gneiting, 2006), and $\boldsymbol{\Sigma}_\mathrm{T}$ is a $n \times n$ temporal ARH(1) variance-covariance matrix. In particular,

$$(\boldsymbol{\Sigma}_\mathrm{S})_{jk} = \frac{1}{2^{\omega-1}\Gamma(\omega)} \left(\frac{s^\mathrm{S}_{jk}}{\nu} \right)^\omega \kappa_\omega \left(\frac{s^\mathrm{S}_{jk}}{\nu} \right), \quad (\boldsymbol{\Sigma}_\mathrm{T})_{jk} = \frac{1}{1-\rho_\varepsilon^2} s^\varepsilon_{jk} \rho_\varepsilon^{d_{jk}},$$

where $\kappa_\omega(\cdot)$ denotes the modified Bessel function of the third kind and order ω, with $\omega > 0$ being a smoothness parameter, $\Gamma(\cdot)$ is the Gamma function, $\nu > 0$ is the scale parameter of the correlation function, and $s^\mathrm{S}_{jk} = \sqrt{(c(k) - c(j))^2 + (r(k) - r(j))^2}$ is the euclidean distance between two plants locations, $(r(k), c(k))$ and $(r(j), c(j))$. As for **Steps 2** and **3**, $s^\varepsilon_{jk} = \sigma^4_\varepsilon h_\varepsilon(t_j) h_\varepsilon(t_k)$, where σ^2_ε is the residual variance, $h_\varepsilon(\cdot)$ is a function of t to the power of 0.8 and ρ_ϵ is the autocorrelation parameter.

At this step, genotypes are assigned to spatial positions/locations following a randomised complete block design, such that each replicate (plant) of a genotype is present in just one block. Depending on the number of replicates (m_g), blocks are accommodated in the row or column direction such that the size of the blocks is the ratio between the number of rows (or columns) and the number of replicates (see Figure 6.1 for an illustrative example of the randomisation). Randomisation is independent for each dataset generated.

Step 5. Calculate M plant trajectories as the sum of the population trajectories, the genotype and plant deviations and the spatio-temporal correlated noise, i.e., $\boldsymbol{y}_i = \boldsymbol{f}_p + \boldsymbol{f}_{g(i)} + \boldsymbol{f}_i + \boldsymbol{\varepsilon}(r(i), c(i), \boldsymbol{t})$.

6.2 Simulation scenarios and set-up

Data is simulated under eight different scenarios given by the combination of the levels of three factors, each with two levels:

1. the between genotype (deviation) variability, $\sigma^2_\mathrm{gen} \in \{\sigma^2_\mathrm{geno.l}, \sigma^2_\mathrm{geno.h}\}$,
2. the between plant (deviation) variability, $\sigma^2_\mathrm{plant} \in \{\sigma^2_\mathrm{plant.l}, \sigma^2_\mathrm{plant.h}\}$, and
3. the number of replicates per genotype, $m_g \in \{3, 10\}$.

For σ^2_gen and σ^2_plant, l stands for low and h for high, that is, $\sigma^2_{\cdot\mathrm{l}} < \sigma^2_{\cdot\mathrm{h}}$. Each scenario is denoted by $(\sigma^2_\mathrm{gen}, \sigma^2_\mathrm{plant}, m_g)$, and they represent different levels of heritability (see, e.g., Rodríguez-Álvarez et al., 2018, for a definition of heritability). For instance, scenarios with $m_g = 10$ replicates per genotype have higher

heritability than those with $m_g = 3$ replicates; and the scenario with the highest heritability is ($\sigma^2_{\text{geno.h}}$, $\sigma^2_{\text{plant.l}}$, 10), that is, the scenario where the between genotype variability is higher than the between plant variability (see Figure 6.3 for a comparison of the heritability among simulation scenarios).

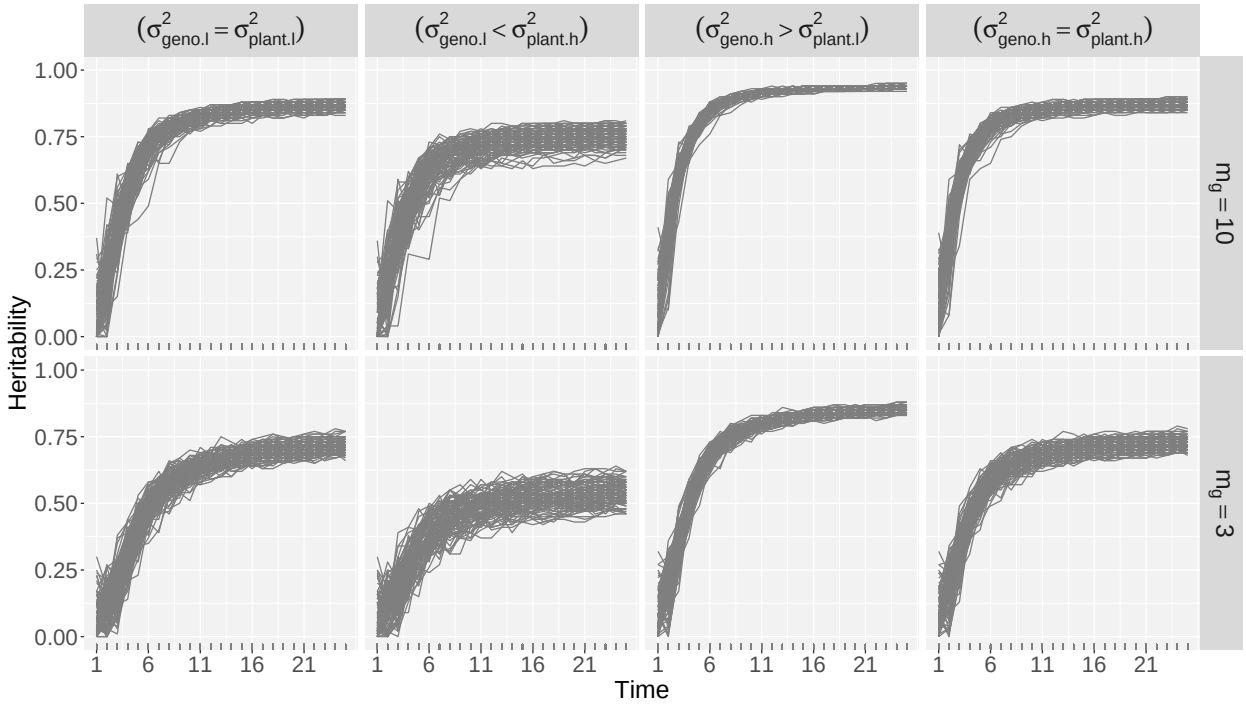


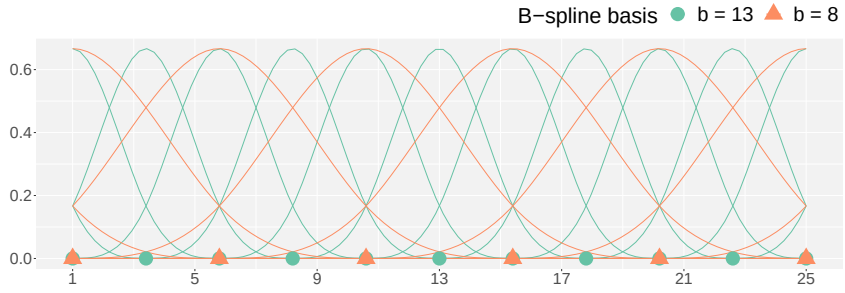
Figure 6.3: Heritability over time for the simulated/true data under eight simulation scenarios (i.e., $(\sigma^2_{\text{geno}}, \sigma^2_{\text{plant}}, m_g)$). Each curve corresponds to one simulated dataset (100 datasets by scenario). Heritability is calculated using `SpATS` (Rodríguez-Álvarez et al., 2018) per time point.

For each scenario, 100 datasets are generated and the simulation settings are described in Table 6.1. We compare the performance of five different configurations for the dimensions of the B-spline bases for the hierarchical components, f_p, f_g and f_i. In particular, we consider $(b_{\text{pop}}, b_{\text{gen}}, b_{\text{plant}}) \in \{(13,13,13),(13,13,8),(13,8,8),(13,9,7),(8,8,8)\}$. The first and fifth configuration aim to assess model's flexibility and over/under-fitting, the second and third configuration evaluate the possible impact of considering different bases dimensions at different levels of the hierarchy, and the fourth configuration assesses model performance under non-nested bases. In this context, nested bases refers to B-spline bases such that the space spanned by $\boldsymbol{B}_{\text{plant}}$ and $\boldsymbol{B}_{\text{gen}}$ are subsets of the space spanned by $\boldsymbol{B}_{\text{pop}}$. For an example see Figure 6.4, and for more technical details we refer to Lee et al. (2013).

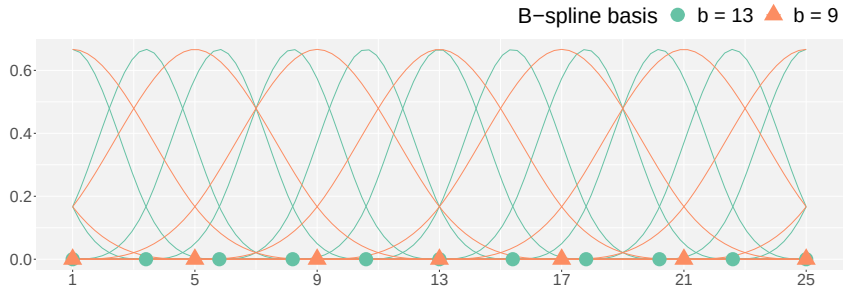
Level	Description	Parameter	Value
Experimental design	Number of populations	p	1
	Number of replicates per genotype	m_g	3 and 10
	Number of time points	n	25
	Number of rows	R	40
	Number of columns	C	21
	Number of genotypes	L	RC/m_g
	Number of plants	M	Lm_g
Population level	Asymptote	a	0.6
	Inflection point	b	12.5
	Growth rate	c	0.4
Genotype level	Between genotypes (deviations) variability	σ^2_{gen}	$\sigma^2_l = 8 \times 10^{-7}$ and $\sigma^2_h = 1.2 \times 10^{-6}$
	Time function	$h(t)$	$8 \times 10^{-1} t^2$
	Autocorrelation	ρ	0.9999
Plant level	Between plants (deviations) variability	σ^2_{plant}	$\sigma^2_l = 8 \times 10^{-7}$ and $\sigma^2_h = 1.2 \times 10^{-6}$
	Time function	$h(t)$	$8 \times 10^{-1} t^2$
	Autocorrelation	ρ	0.9999
Spatio-temporal correlated noise	Residual variance	σ^2_ε	1×10^{-8}
	Time function	$h_\varepsilon(t)$	$1000 \times t^{0.8}$
	Autocorrelation	ρ_ε	0.5
	Smoothness parameter	ω	10000
	Scale parameter	ν	5

Table 6.1: Simulation settings. Reference values are based on the leaf area (m^2 plant^{-1}) data from the PhenoArch platform. For this data, the maximum leaf area is approximately 0.5 (we fixed the asymptote for the population trajectory at $a = 0.6$), the average growth rate is 0.01 (we increased this value to $c = 0.4$ to obtain S-shape curves), the between plants variability at the beginning of the experiment is 4.5×10^{-7} (we set $\sigma^2_l = 8 \times 10^{-7}$ and $\sigma^2_h = 1.2 \times 10^{-6}$), and the maximum first-order autocorrelation for observations of plant trajectories is 0.9986 (we used $\rho = 0.9999$ to obtain smoother curves).

In addition, to make the one and two-stage approaches comparable, we set the dimensions of the B-spline bases of the model components that are common to the two approaches to the same values. Consequently,



(a) Nested B−spline bases



(b) Non−nested B−spline bases

Figure 6.4: Illustrative visualisation of (a) nested (e.g., $(b_1 = 13, b_2 = 8)$) and (b) non-nested B-spline bases (e.g., $(b_1 = 13, b_2 = 9)$).

- Two-stage approach
  - First stage: we fit a SpATS model separately for each time point, with $b_2 = b_3 = 13$ B-splines for the two-dimensional smooth function, h_S (see also (4.1) and (3.26)).
  - Second stage: we fit the psHDM (4.6), using the five different configurations, $(b_{\text{pop}}, b_{\text{gen}}, b_{\text{plant}})$, previously described for the hierarchical functions at population, f_p, genotype, f_g, and plant levels f_i.
- One-stage approach: we fit the spatio-temporal psHDM (5.2) with $b_{\text{row}} = b_{\text{col}} = 13$ B-splines for the random row, f_r, and column, f_c, effects; $b_1 = b_2 = b_3 = 13$ for the spatio-temporal smooth function, f_{ST}; and the five different configurations, $(b_{\text{pop}}, b_{\text{gen}}, b_{\text{plant}})$, previously described for the hierarchical functions at population, f_p, genotype, f_g, and plant levels f_i.

We note that the number of simulated datasets (100 in total) are limited by the running time of the one-stage approach (which is the most time and memory consuming). Computations were performed in the the BCAM HPC (Basque Center for Applied Mathematics High Performance Computing), with three types of nodes ((i). 4 nodes (1 with Tesla GPU), Processor Intel$^{(R)}$ Xeon$^{(R)}$ CPU E5-2680 v3, 24 core - 128GB memory, (ii). 12 nodes, Processor Intel$^{(R)}$ Xeon$^{(R)}$ CPU E5-2683 v4, 32 core - 256GB memory, and (iii). 2 nodes, Processor Intel$^{(R)}$ Xeon$^{(R)}$ Gold 6140 CPU, 72 core - 384GB memory), and a (64-bit) R 4.0.4-foss-2020b. Running times (d-h:m:s) for 100 datasets with the one-stage approach, one simulation scenario, and

one B-spline basis configuration, ranged between $2-15:58:19$ and $19-05:33:23$. Similarly, memory consumption ranged between 6 and 17GB approximately. Our algorithms did not present any convergence problem (for either approach). It should be noted that at the time of the simulations the LMMsolver was not implemented in our codes, which leads to longer calculation times.

6.3 Simulation results

We use the logarithm of the root mean square error (log(RMSE)) as a performance measure to compare the simulated and the estimated curves. Thus, the lower the value, the better the performance. This measure is calculated for the combination of the five basis configurations, the two models, and the eight simulation scenarios. In what follows, we present the results of the simulation study by hierarchy level (populations, genotypes and plants). General conclusions are given in the final remarks.

6.3.1 Population-level results

For curves at the population level (trajectories and first-order derivatives) we compare the simulated population trajectory, $\boldsymbol{f}_p$ (which is fixed for all simulations), and the estimated population trajectory, $\boldsymbol{f}_p^{(\gamma)} = (f_p^{(\gamma)}(t_1), \ldots, f_p^{(\gamma)}(t_n))^T$, for each simulated dataset ($\gamma = 1, \ldots, 100$) as follows (for trajectories, as illustration, but the same idea is followed for their first-order derivatives)

$$\text{RMSE}_\gamma = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(f_p(t_j) - \hat{f}_p^{(\gamma)}(t_j)\right)^2}.$$

In Figure 6.5, we use boxplots to compare the results using the two approaches (the one-stage in continuous borders and the two-stage in dotted borders) for the eight simulation scenarios and the five B-spline basis configurations (in colours). For all scenarios with the same number of replicates, results are similar when comparing the B-spline basis configurations. Furthermore, for scenarios with $m_g = 10$ replicates, the one-stage approach consistently outperforms the two-stage approach for all B-spline basis configurations. When the non-nested basis configuration (i.e., $b_{\text{pop}} = 13$, $b_{\text{gen}} = 9$, and $b_{\text{plant}} = 7$, in blue) is used, the two-stage performs the worst for scenarios with $m_g = 10$ replicates. We found that results for scenarios with $m_g = 3$ replicates and the B-spline basis configuration (8,8,8), in purple, performs the worst for the one- and two-stage approaches.

For a more in-depth analysis of the population trajectories, we plot in Figure 6.6 the simulated, $\boldsymbol{f}_p$ in black, and estimated, $\hat{\boldsymbol{f}}_p$ in grey, population trajectories when using the two approaches with the non-nested B-spline basis configuration, for the eight simulation scenarios. We observe that for all scenarios with $m_g = 10$ replicates and the two-stage approach, some of the simulated population trajectories have

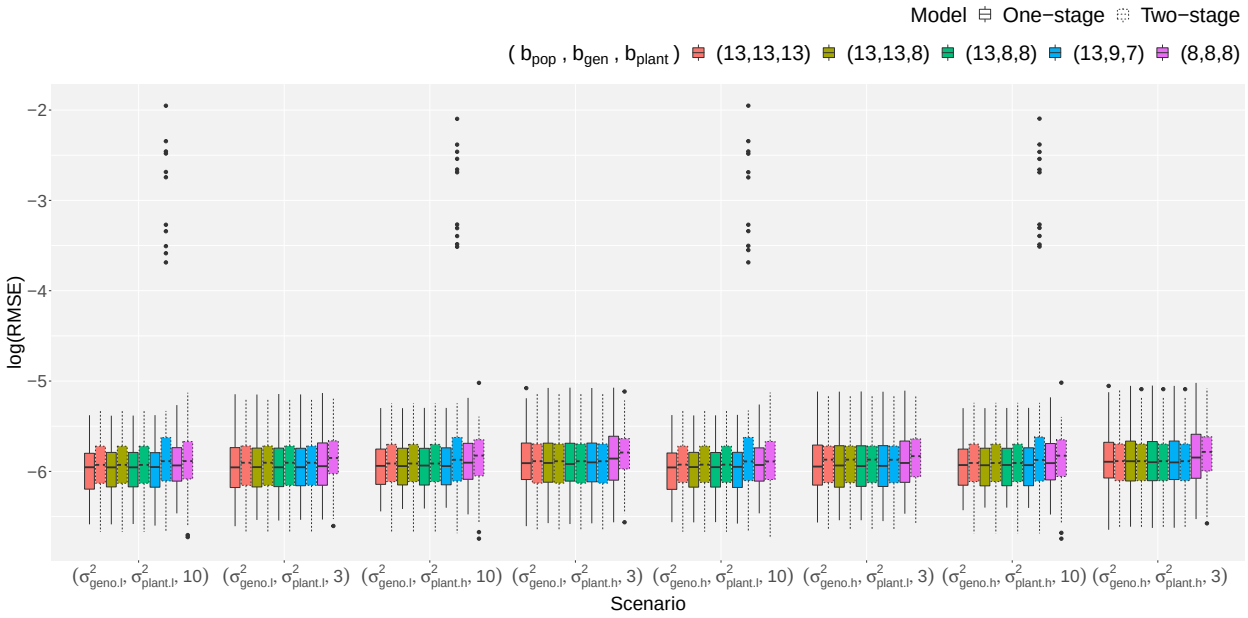


Figure 6.5: Simulation results: comparison of the simulated/true ($\boldsymbol{f}_p$) and estimated ($\hat{\boldsymbol{f}}_p$) population trajectories, for eight simulation scenarios ($\sigma^2_{\text{geno}}, \sigma^2_{\text{plant}}, m_g$), using the one- and two-stage approaches, and the five B-spline basis configurations (b_{pop},b_{gen},b_{plant}) at population, genotype and plant level, respectively.

an unexpected and wild behaviour, which is more evident in the first time period (i.e., before $t = 12$, that corresponds with the inflection point, $b = 12.5$, as indicated in Table 6.1, used to simulate the population trajectory).

In Figure 6.7, we use boxplots to compare the log(RMSE) calculated for the first-order derivative curves using the two approaches (the one-stage in continuous borders and the two-stage in dotted borders) for the eight simulation scenarios and the five B-spline basis configurations (in colours). For these first-order derivative curves, the one-stage approach consistently outperforms the two-stage approach. Results are stable among simulation scenarios with the same number of replicates. The biggest difference between scenarios with $m_g = 10$ and $m_g = 3$ replicates is that the non-nested B-spline basis configuration (i.e., $b_{\text{pop}} = 13$, $b_{\text{gen}} = 9$, and $b_{\text{plant}} = 7$, in blue) is the one that performs the worst for the two-stage approach. Opposite to what we found for the population trajectories (see Figure 6.5), for all simulation scenarios, the B-spline basis configuration (8,8,8) in purple performs the best for the one- and two-stage approaches.

In Figure 6.8, we use as illustration one simulation scenario (the one with the highest heritability, $(\sigma^2_{\text{geno.h}}, \sigma^2_{\text{plant.l}}, 10)$, see Figure 6.3) to show the simulated, $\boldsymbol{f}'_p$ in black, and the estimated, $\hat{\boldsymbol{f}}'_p$ in grey, first-order derivative of the population trajectories for the five B-spline basis configurations using the two approaches. Three things are noteworthy: firstly, as a consequence of the unexpected and wild behaviour of the estimated population trajectories with the non-nested B-spline basis configuration, their first-order derivative behaves similarly; secondly, we observe a rougher behaviour and more variation at the bound-

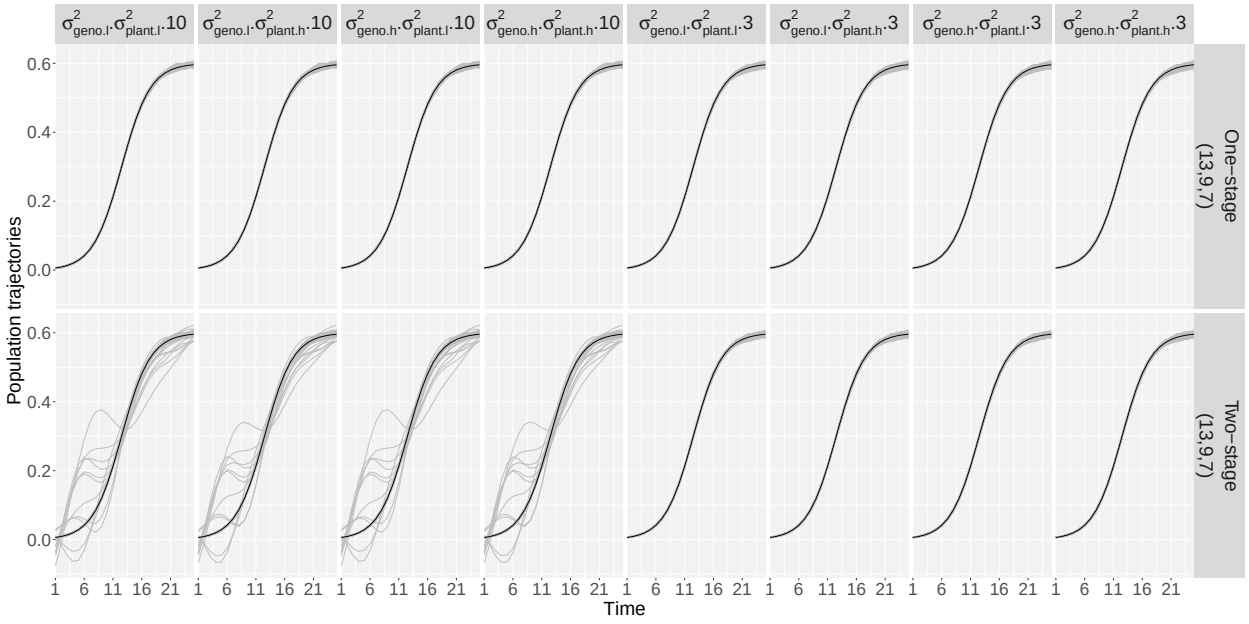


Figure 6.6: Simulation results: comparison of the simulated/true ($\boldsymbol{f}_p$, in black) and estimated ($\hat{\boldsymbol{f}}_p$, in grey) population trajectories when a non-nested cubic B-spline basis configuration is used (i.e., $b_{\text{pop}} = 13$, $b_{\text{gen}} = 9$, and $b_{\text{plant}} = 7$), for the eight simulation scenarios ($\sigma^2_{\text{geno}}, \sigma^2_{\text{plant}}, m_g$), and the one- and two-stage approaches.

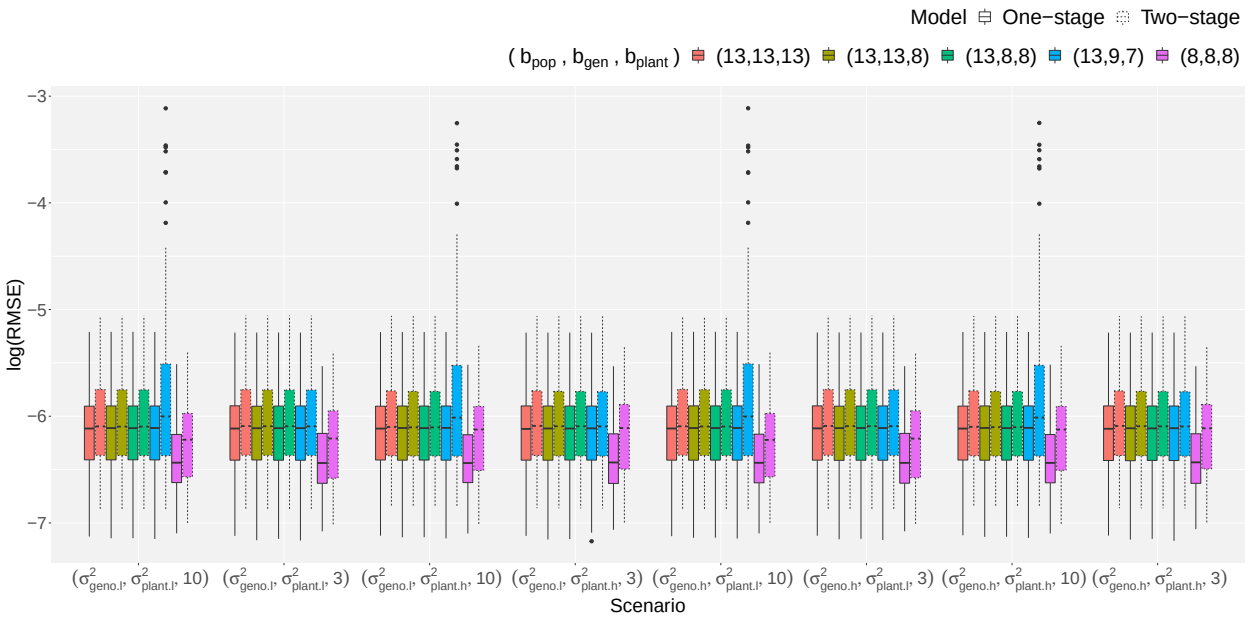


Figure 6.7: Simulation results: comparison of the simulated/true ($\boldsymbol{f}'_p$) and estimated ($\hat{\boldsymbol{f}}'_p$) first-order derivative of the population trajectories, for eight simulation scenarios ($\sigma^2_{\text{geno}}, \sigma^2_{\text{plant}}, m_g$), using the one- and two-stage approaches, and the five B-spline basis configurations (b_{pop},b_{gen},b_{plant}) at population, genotype and plant level, respectively.

aries (especially on the right side); and lastly, estimated first-order derivative curves for the (8, 8, 8) B-spline basis configuration are smoother (less wiggly) and with less variation at the boundaries, and consequentely with better performance for the two-approaches.

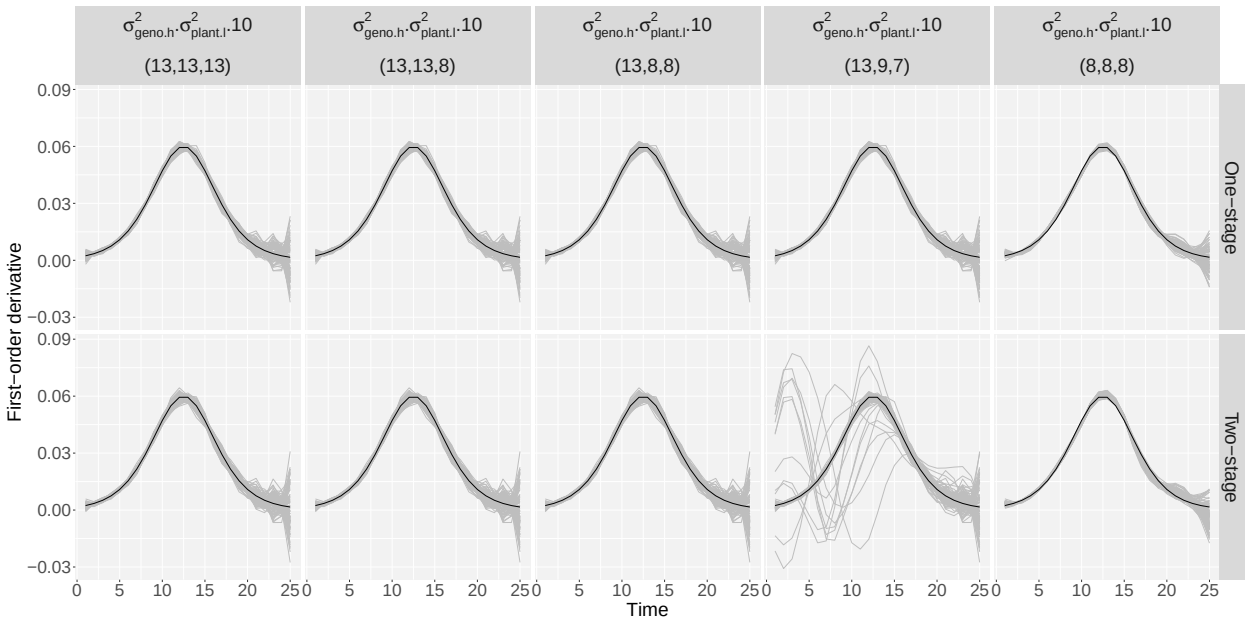


Figure 6.8: Simulation results: comparison of the simulated/true ($\boldsymbol{f}'_p$, in black) and estimated ($\hat{\boldsymbol{f}}'_p$, in grey) first-order derivative of the population trajectories for one simulation scenario (($\sigma^2_{\text{geno.h}}, \sigma^2_{\text{plant.l}}, 10$), as illustration), using the one- and two-stage approaches, and the five B-spline basis configurations (b_{pop},b_{gen},b_{plant}) at population, genotype and plant level, respectively.

6.3.2 Genotype-level results

We now focus on the results at the genotype level, which is the decision-making level for plant breeders. Our main question is: Can we properly reproduce the genotype deviations? For these curves (trajectories and deviations) and for each genotype (genotypic curves are the same for all simulations , we calculate (for deviations, as illustration, but the same applies to trajectories)

$$\text{RMSE}_{g,\gamma} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(f_g(t_j) - \hat{f}_g^{(\gamma)}(t_j)\right)^2}.$$

Boxplots in Figure 6.9 depict the log(RMSE) when comparing the simulated and estimated genotype deviations for the eight simulation scenarios, colours represent different B-spline basis configurations, and line type box border stands for the approach used (continuous border for the one-stage and dotted border for the two-stage). Differences among simulation scenarios are more remarkable at this hierarchy level than at

the population level. We assume these differences are related to heritability (see Figure 6.3). For instance, scenarios with $m_g = 10$ replicates have higher heritability and better performance when compared with scenarios with $m_g = 3$ replicates; and those scenarios with $m_g = 10$ replicates, and $(\sigma^2_{\text{geno.l}}, \sigma^2_{\text{plant.l}})$ and $(\sigma^2_{\text{geno.h}}, \sigma^2_{\text{plant.l}})$ show the best performance (as well as the highest heritability). For this hierarchy level, we consistently observe (as for the population trajectories and their first-order derivatives in Figures 6.7 and 6.8) the fitting problem when the non-nested B-spline basis configuration is used with the two-stage approach (i.e., $b_{\text{pop}} = 13$, $b_{\text{gen}} = 9$, and $b_{\text{plant}} = 7$, in blue), for scenarios with $m_g = 10$ replicates. Opposite to what we found at the population level, results for the one- and two-stage approach and the different B-spline basis configurations (except for $(13, 9, 7)$) are very similar. A very slight difference can be noticed for the $(13, 13, 8)$ B-spline basis configuration (in olive), where the one-stage approach performs the "worst"; we note that in this case, we are using different basis dimensions at genotype and plant levels.

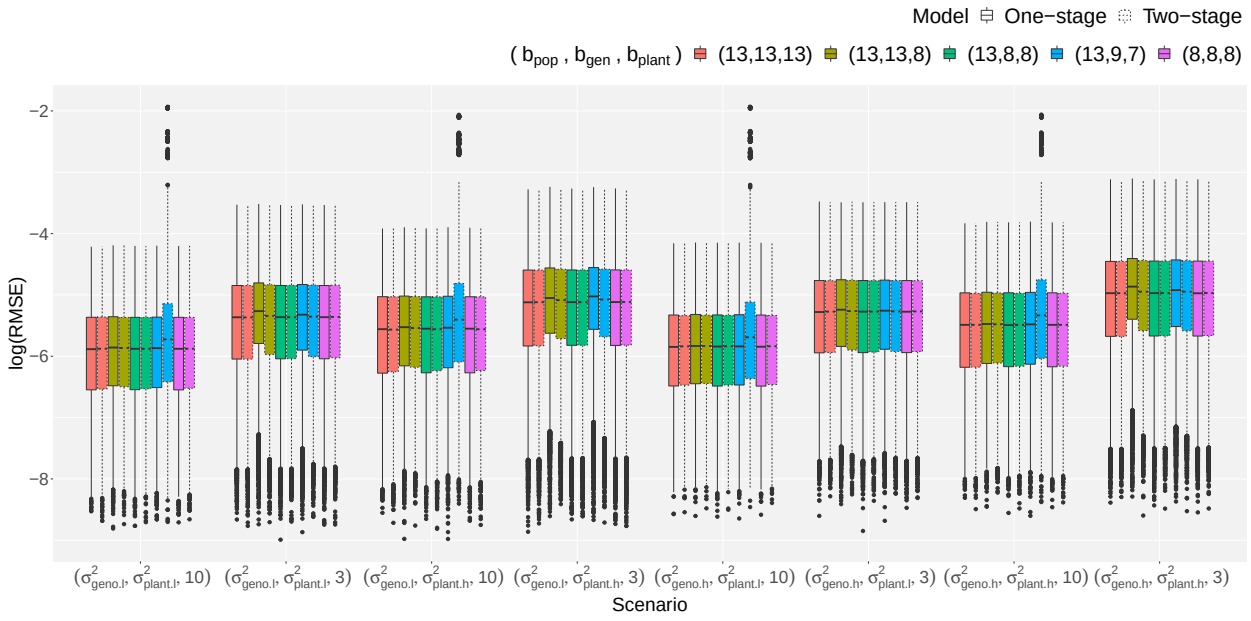


Figure 6.9: Simulation results: comparison of the simulated/true ($\boldsymbol{f}_g$) and estimated ($\hat{\boldsymbol{f}}_g$) genotype deviations, for eight simulation scenarios $(\sigma^2_{\text{geno}}, \sigma^2_{\text{plant}}, m_g)$, using the one- and two-stage approaches, and the five B-spline basis configurations (b_{pop},b_{gen},b_{plant}) at population, genotype and plant level, respectively.

To go further in the results shown in Figure 6.9, we depict in Figure 6.10 for three genotypes ($g = 5, 10, 72$), one simulation scenario ($(\sigma^2_{\text{geno.h}}, \sigma^2_{\text{plant.l}}, 10)$, the one with the highest heritability), and the non-nested B-spline basis configuration $(13, 9, 7)$, the simulated/true (in black) and estimated (red for the one-stage and blue for the two-stage approaches) genotype deviations; red and blue lines are pointwise averages of estimated curves, and shaded areas are bands constructed using the pointwise 2.5% and 97.5% quantiles across simulations. As for the population level in Figure 6.6, we observe the unexpected behaviour for some of the simulated genotype deviations when the non-nested, $(13, 9, 7)$, B-spline basis configuration is used,

being the inflection point (around $b = 12.5$) a point of major change in the shape of the deviation.

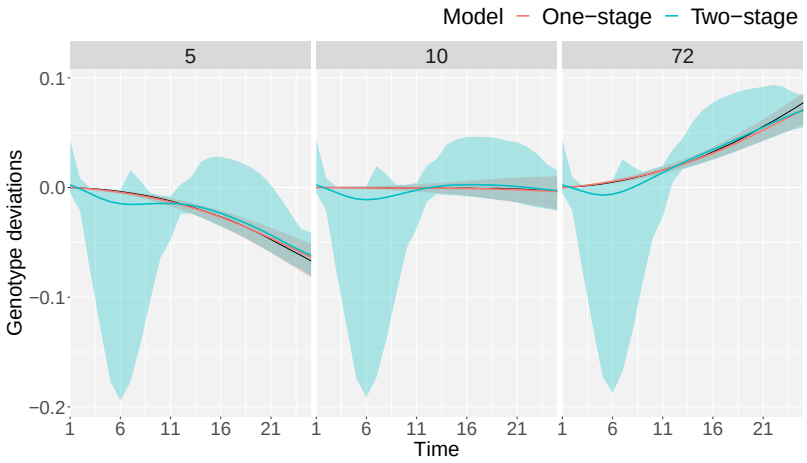


Figure 6.10: Simulation results: comparison of the simulated/true ($\boldsymbol{f}_g$, in black) and pointwise average of estimated genotype deviations ($\hat{\boldsymbol{f}}_g$) calculated with the one-stage (in red) and two-stage (in blue) for three genotypes ($g = 5, 10, 72$, as illustration), for one simulation scenario ($\sigma^2_{\text{geno.h}}, \sigma^2_{\text{plant.l}}, 10$) and the non-nested B-spline basis configuration ($b_{\text{pop}} = 13$, $b_{\text{gen}} = 9$, $b_{\text{plant}} = 7$). The shaded area are bands constructed using the pointwise 2.5% and 97.5% quantiles across simulations.

In Figure 6.11, we zoom in on the performance results with the two approaches for the simulation scenario ($\sigma^2_{\text{geno.h}}, \sigma^2_{\text{plant.l}}, 10$), when omitting the $(13, 9, 7)$ basis configuration. We plot the genotype deviations for three genotypes ($g = 5, 10, 72$). In all three cases, the shaded area (i.e., the bands constructed using the pointwise 2.5% and 97.5% simulation quantiles) contains the simulated/true genotype deviation, and the pointwise average of estimated curves for the one- and two-stage approaches are very close between them (red and blue lines), as well as to the simulated/true genotype deviation (in black).

We finally present, in Figure 6.12, some extra results for the simulation scenario with the lowest heritability (i.e., ($\sigma^2_{\text{geno.l}}, \sigma^2_{\text{plant.h}}, 3$), see Figure 6.3). Recall that scenarios with $m_g = 3$ replicates are not strongly sensitive to the non-nested basis configuration as they are the scenarios with $m_g = 10$ replicates. We observe an additional result here: the estimated genotype deviations with the B-spline basis configurations $(13, 13, 8)$ and $(13, 9, 7)$ seem to be linear, even when smooth (non-linear) terms are included in the model specification (5.7). We find the shaded areas (i.e., the bands constructed using the pointwise 2.5% and 97.5% simulation quantiles) for this scenario to be slightly wider than those for the scenario shown in Figure 6.11, i.e., for ($\sigma^2_{\text{geno.h}}, \sigma^2_{\text{plant.l}}, 10$), which makes sense as there are fewer replicates.

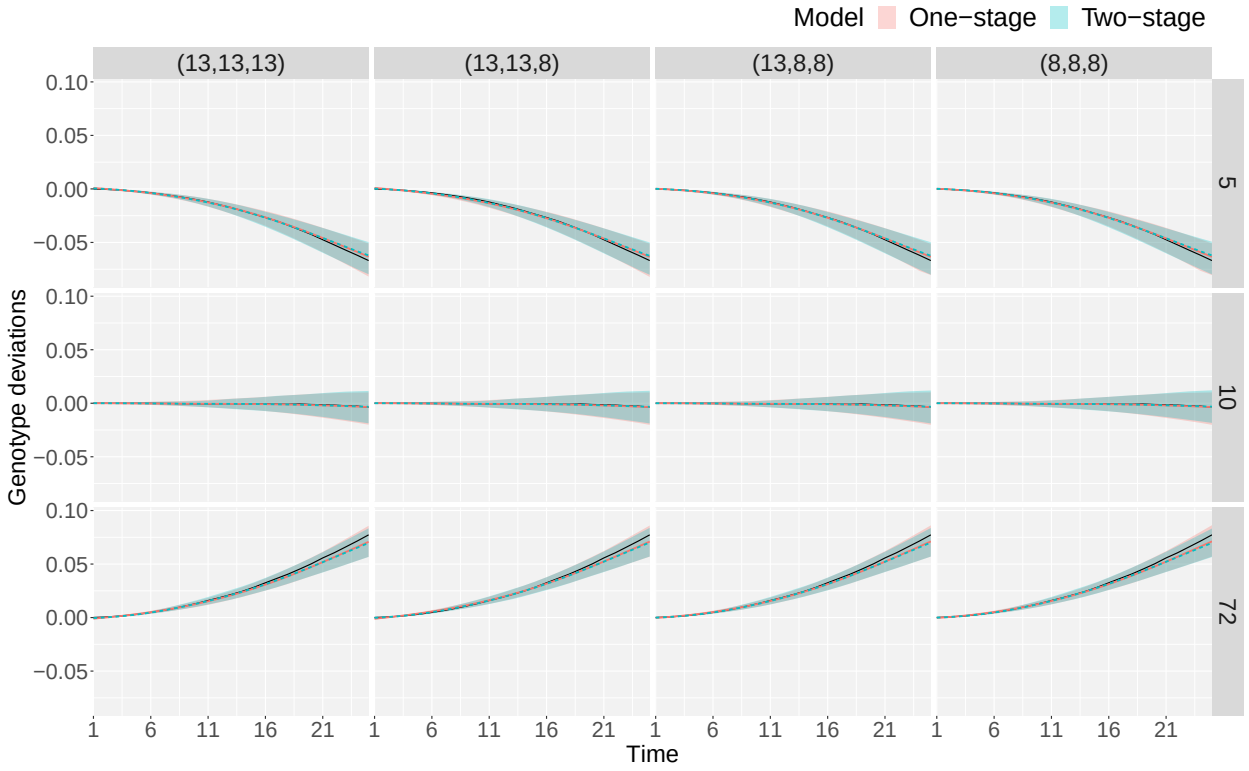


Figure 6.11: Simulation results: comparison of the simulated/true ($\boldsymbol{f}_g$, in black) and pointwise average of estimated ($\hat{\boldsymbol{f}}_g$) genotype deviations for three genotypes ($g = 5, 10, 72$, as illustration), for one simulation scenario ($\sigma^2_{\text{geno.h}}, \sigma^2_{\text{plant.l}}, 10$), using the one- (in red) and two-stage (in blue) approaches, and four (excluding the non-nested basis configuration) B-spline basis configurations (b_{pop},b_{gen},b_{plant}) at population, genotype and plant level, respectively. The shaded area are bands constructed using the pointwise 2.5% and 97.5% quantiles across simulations.

6.3.3 Plant-level results

Finally, for curves at the plant level (trajectories and deviations), and for each plant we have (for deviations, as illustration, but it can also be calculated for trajectories)

$$\text{RMSE}_{i,\gamma} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(f_i^{(\gamma)}(t_j) - \hat{f}_i^{(\gamma)}(t_j)\right)^2}$$

Figure 6.13 shows the boxplots used to compare the performance of the estimated plant trajectories for the eight simulation scenarios when using the one-stage (in continuous box borders) and the two-stage (in dotted box borders) approaches and the five B-spline basis configurations (in colours). As for the results at the genotype level in Figure 6.9, we find differences in results among simulation scenarios, being ($\sigma^2_{\text{geno.h}}, \sigma^2_{\text{plant.l}}, 10$) and ($\sigma^2_{\text{geno.l}}, \sigma^2_{\text{plant.l}}, 10$), the two scenarios with the best performance (and the ones with

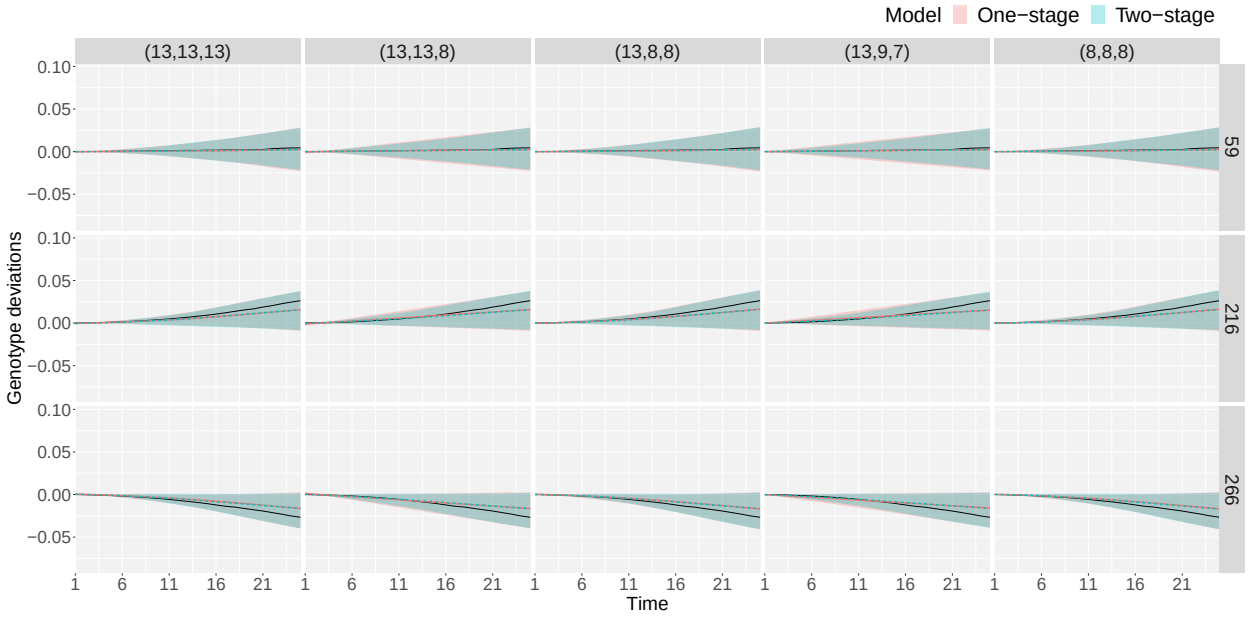


Figure 6.12: Simulation results: comparison of the simulated/true ($\boldsymbol{f}_g$, in black) and and pointwise average of estimated ($\hat{\boldsymbol{f}}_g$) genotype deviations for three genotypes ($g = 59, 216, 266$, as illustration), for one simulation scenario ($\sigma^2_{\text{geno.l}}, \sigma^2_{\text{plant.h}}, 3$), using the one- (in red) and two-stage (in blue) approaches, and the five B-spline basis configurations (b_{pop},b_{gen},b_{plant}) at population, genotype and plant level, respectively. The shaded area are bands constructed using the pointwise 2.5% and 97.5% quantiles across simulations.

the highest heritability, see Figure 6.3); and ($\sigma^2_{\text{geno.l}}, \sigma^2_{\text{plant.h}}, 3$) and ($\sigma^2_{\text{geno.h}}, \sigma^2_{\text{plant.h}}, 3$), the two scenarios with the worst performance (and the ones with the lowest heritability). For the two scenarios with the highest heritability, we also observe bigger differences between the one- and two-stage approaches for the five B-spline basis configurations, and the one-stage consistently outperforms the two-stage approach. In contrast to the results at population (see Figures 6.5 and 6.7) and genotype (see Figure 6.9) levels, at the plant level, we do not find the strange and wild behaviour of the non-nested B-spline basis for none of the scenarios studied.

At the plant level, we are also interested in the fitted values, i.e., the plant trajectories, which are the curves the breeders finally observe and measure. In Figure 6.14, we compare the simulated (thicker, transparent lines) and the estimated plant trajectories (one-stage in continuous lines and two-stage in dotted lines) for the plants (in colours) in three genotypes ($g = 90, 160, 224$, as illustration) for one simulation scenario ($\sigma^2_{\text{geno.h}}, \sigma^2_{\text{plant.l}}, 3$) and one simulated dataset, and the five B-spline basis configurations. The difference between Figures 6.14(a) and 6.14(b) is the simulated data that is used to compare with the estimated plant curves, $\hat{\boldsymbol{f}}_p + \hat{\boldsymbol{f}}_g + \hat{\boldsymbol{f}}_i$. In Figure 6.14(a), we use the simulated plant trajectories when including the spatio-temporal correlated noise, $\varepsilon(r, c, \boldsymbol{t})$, while in Figure 6.14(b), we omit this term from the simulation model (6.1). We consistently observe slight differences between the one- and two-stage approaches, which

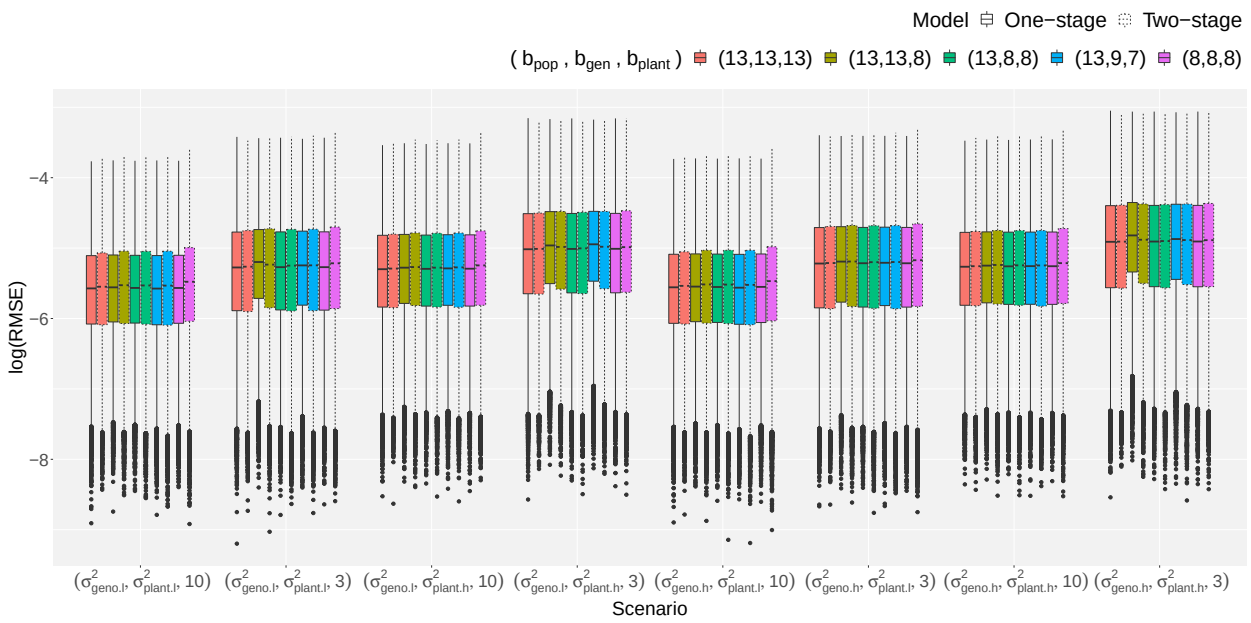


Figure 6.13: Simulation results: comparison of the simulated/true ($\boldsymbol{f}_i$) and estimated ($\hat{\boldsymbol{f}}_i$) plant deviations, for eight simulation scenarios ($\sigma^2_{\text{geno}}, \sigma^2_{\text{plant}}, m_g$), using the one- and two-stage approaches, and the five B-spline basis configurations (b_{pop},b_{gen},b_{plant}) at population, genotype and plant level, respectively.

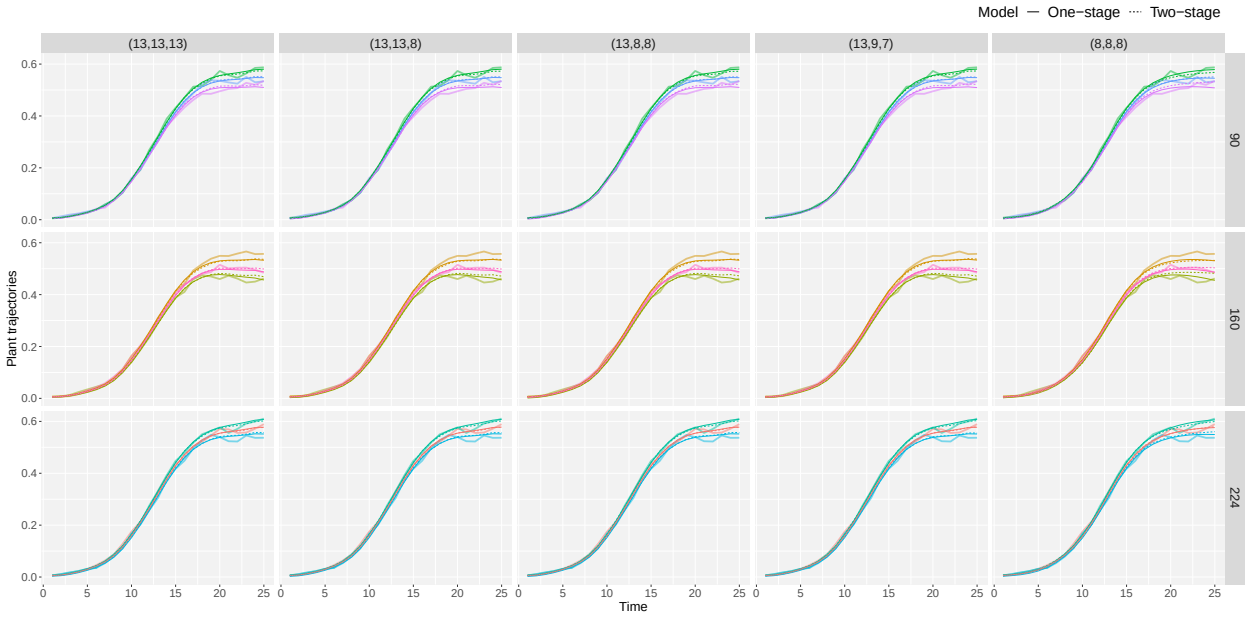
properly fit the simulated plant curves in Figure 6.14(b).

As for the genotype level (see Figures 6.10, 6.11, and 6.12), we are also interested in plant deviations. Figure 6.15 depicts the simulated (thicker, transparent lines) and estimated (continuous lines for the one-stage and dotted lines for the two-stage) plant deviations (colours represent different plants) for three genotypes ($g = 90, 160, 224$, as illustration) in one simulation scenario ($\sigma^2_{\text{geno.h}}, \sigma^2_{\text{plant.l}}, 3$), for one simulated dataset and the five B-spline basis configurations. Genotype deviations are also depicted as reference curves (simulated genotype deviations in black and estimated genotype deviations using the one-stage in continuous red lines and the two-stage in continuous blue lines). While we observe slight differences between the one- and two-stage approaches for the plant trajectories in Figure 6.14, they are noteworthy for plant deviations.

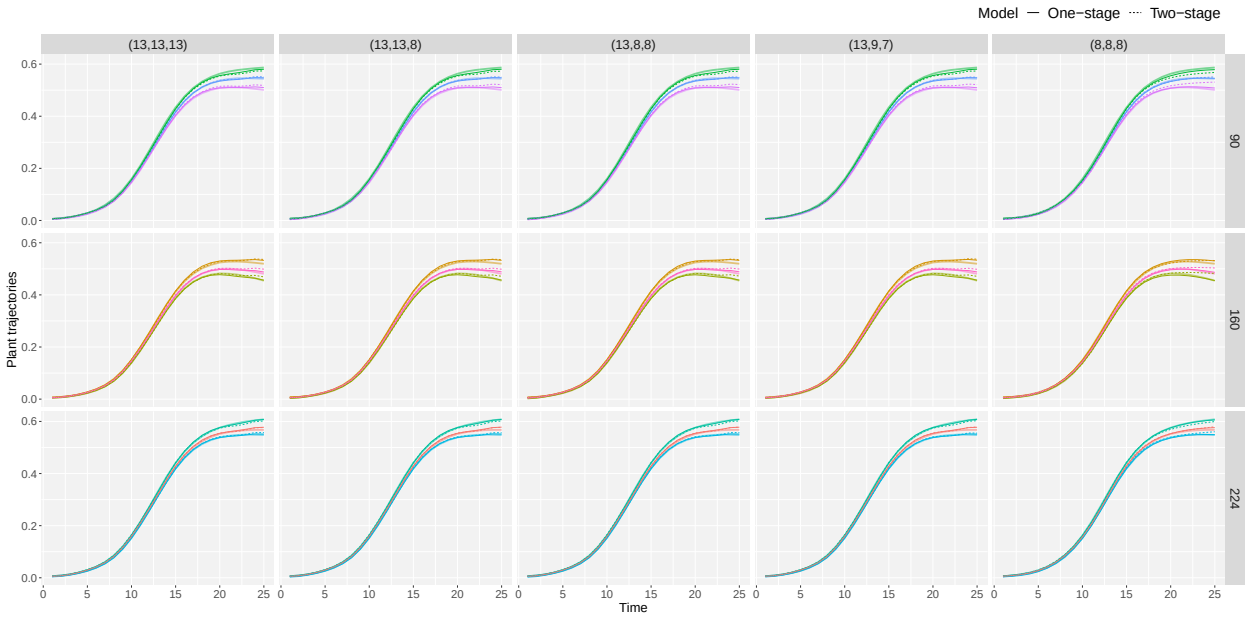
6.3.4 Final remarks

We close this chapter with a summary of the main findings of the simulation study. We mainly refer to the results in Figures 6.5, 6.7, 6.9 and 6.13

- Results show more differences in the performance among scenarios at the genotype and plant levels, where scenarios with the highest heritability performed better.



(a) Comparison with simulated plant trajectories, $\boldsymbol{f}_p + \boldsymbol{f}_g + \boldsymbol{f}_i + \varepsilon(r, c, \boldsymbol{t})$



(b) Comparison with spatially corrected plant trajectories, $\boldsymbol{f}_p + \boldsymbol{f}_g + \boldsymbol{f}_i$

Figure 6.14: Simulation results: comparison of the (**a**) simulated/true plant trajectories, $\boldsymbol{f}_p + \boldsymbol{f}_g + \boldsymbol{f}_i + \varepsilon(r, c, \boldsymbol{t})$ (thicker, transparent lines) and (**b**) spatially corrected simulated plant trajectories, $\boldsymbol{f}_p + \boldsymbol{f}_g + \boldsymbol{f}_i + \boldsymbol{\varepsilon}_i$ (thicker, transparent lines) with the estimated, $\hat{\boldsymbol{f}}_p + \hat{\boldsymbol{f}}_g + \hat{\boldsymbol{f}}_i$, plant trajectories for the plants (in colours) in three genotypes ($g = 90, 160, 224$, as illustration), for one simulation scenario ($\sigma^2_{\text{geno.h}}, \sigma^2_{\text{plant.l}}, 3$) and one simulated dataset, using the one-stage (continuous lines) and two-stage (dotted lines) approaches, and the five B-spline basis configurations (b_{pop},b_{gen},b_{plant}) at population, genotype and plant levels, respectively.

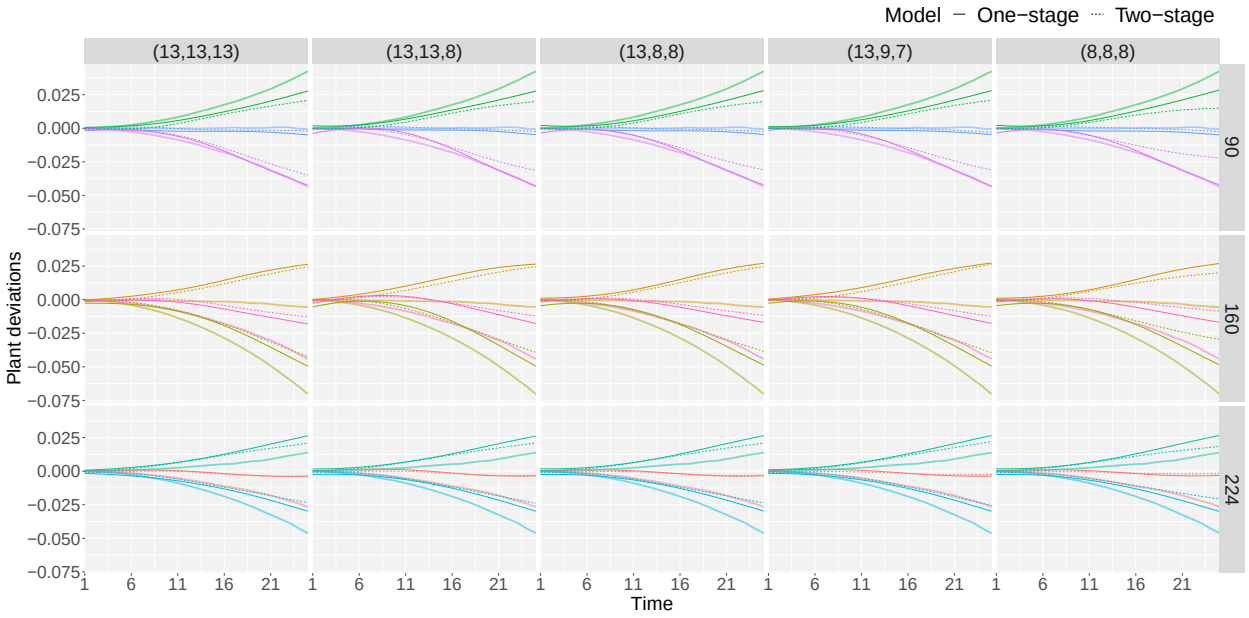


Figure 6.15: Simulation results: comparison of the simulated/true plant deviations, $\boldsymbol{f}_i$ (thicker, transparent lines), and estimated plant deviations, $\hat{\boldsymbol{f}}_i$, for plants (in colours) in three genotypes ($g = 90, 160, 224$, as illustration we show the same used in Figure 6.14), for one simulation scenario ($\sigma^2_{\text{geno.h}}, \sigma^2_{\text{plant.l}}, 3$) and one simulated dataset, using the one-stage (in continuous lines) and two-stage (in dotted lines) approaches, and the five B-spline basis configurations (b_{pop},b_{gen},b_{plant}) at population, genotype and plant level, respectively.

- We found slight differences in the performance of the one- and two-stage approaches at the genotype and plant levels. The largest differences are for the population trajectories (for scenarios with $m_g = 10$ replicates) and their first-order derivatives. When differences are found, the one-stage consistently outperforms the two-stage approach.

- Similar results were obtained with different B-spline basis configurations. Nevertheless, we found the non-nested B-spline basis configuration, $(13, 9, 7)$, problematic (unexpected and wild behaviour) at the population (for trajectories and their first-order derivatives) and genotype levels for scenarios with $m_g = 10$ replicates. We are also struck by the ambivalent result for the B-spline basis configuration $(8, 8, 8)$: it performs the best for the first-order derivative, i.e., smoother and with less variation at the boundaries. However, it does not show good behaviour when the population trajectories and plant deviations are estimated.

- We have proposed two P-spline-based approaches (in one and two stages), which means that as the number of plants and basis dimensions increase, so it does the computation time (scalability problem). These simulation results, jointly with data analyses performed during the research period, have shown that results may be sensitive (and in some cases unreliable) to using different bases dimensions. We, therefore, recommend choosing the same number of B-spline basis functions for the three levels of the

hierarchy (i.e., b_{pop}, b_{gen}, and b_{plant}), as well as for the row and column random effects (i.e., b_{row} and b_{col}), even if this increases computation. Regarding the number of B-spline basis functions used for the three-dimensional surface (in the row, column and time directions, i.e., b_1, b_2, and b_3 in the one-stage approach), we suggest keeping them relatively small to enable the solution to run on standard computers. We expect this choice allows to keep the number of coefficients at a reasonable level (i.e., a trade-off between flexibility and dimensionality).

Chapter 7

Data application: HTP data analysis

The aim of this chapter is twofold. We present the results by modelling the two HTP datasets presented in Chapter 2 using both the one-stage (Chapter 5) and two-stage (Chapter 4) approaches, while simultaneously comparing them. For that purpose, we set the dimensions of the B-spline basis that are common to both approaches to the same value, which are limited by running times for the one-stage approach. For each dataset, we first present the models configuration used, then we show the most relevant results and compare the one- and two-stage approaches. Finally, we extract some time-independent attributes to characterise genotypes. For the FIP data, we additionally comment on the genotype consistency across the three trials. Computations were performed in a (64-bit) `R` 4.2.1 and a 1.60GHz Dual-Core^{TM} i5 processor computer with 16GB of RAM and macOS Monterrey Version 12.5. Chapter 8 shows the functionalities of the code implemented for this thesis. More specifically, to obtain the results for the two-stage approach we use the code available in the `R`-package `statgenHTP` (Millet et al., 2022), and for the results with the one-stage approach the codes are available in https://gitlab.bcamath.org/dperez/htp_one_stage_approach.

7.1 PhenoArch results

A data description of the PhenoArch platform was presented in Section 2.1. We recall that this dataset consists of $n = 32$ leaf area measurements on $M = 1656$ plants (in a grid of R rows $\times$ C columns = 60×28), where $L = 90$ genotypes were tested, grouped in two panels (60 genotypes in Panel 1 and 30 genotypes in Panel 2). All genotypes were tested under two levels of soil water content (WW and WD). For a proper analysis of this dataset, the factorial structure of panels and treatments (crossed effects) should have been included in our model. For simplicity, we treated the combinations of panels and treatments as a single factor with 4 levels ("populations" with $K = 4$ levels, that is, Panel 1 - WD, Panel 1 - WW, Panel 2 - WD and Panel 2 - WW) and $L = 180$ "genotypes" (60 genotypes in Panel 1 and WW treatment, 60 genotypes in Panel 1 and WD treatment, 30 genotypes in Panel 2 and WW treatment and 30 genotypes in Panel 2 and WD treatment). As such, in our analysis, the "population" effects should be understood as the effects resulting from the panel-by-water regime combination, while the "genotypic" effects should be interpreted as the effects arising from the genotype-by-water regime combination. Additionally, we eliminated the last time point (i.e., $n = 31$), and plants with 20 or less measurements (i.e., $M = 1648$).

7.1.1 PhenoArch results: Approaches specification

In this section we comment on the approaches specification, and more precisely on the number of coefficients and variance parameters to give the reader an idea of the complexity of the models used. For the two-stage approach, we fitted a SpATS model (see (4.2)) for each individual measurement time point of the leaf area data. In addition to the spatial trend $h_S(r, c)$, and the genotypic effects h_g, the model included the population (panel by water regime combination) as fixed effect h_p, and the row and column positions as random effects, h_r and h_c. We illustrate the difference in the fitted values when modelling genotypic effects h_g as random (BLUPs) or fixed (BLUEs) in Section 7.1.2. A different genetic variance for each of the four populations (panel by water regime combination) was considered when genotypes were modelled as random. Regarding the spatial trend (i.e., the bi-dimensional tensor-product P-spline), B-spline bases of dimension $b_2 = b_3 = 8$ were chosen in the row and column directions. Under this configuration, the mixed model formulation (4.2) of the SpATS model (for one time point) when modelling genotypes as random has $(K + L + R + C + (b_2 b_3 - 1) = 4 + 180 + 60 + 28 + (8(8) - 1) =)$ 335 coefficients and 12 variance parameters: four genotype variances associated with each population; two variances for the random row and column effects; five variances for the smooth spatial function $h_s(r, c)$, where we assume the PS-ANOVA decomposition in the variance-covariance matrix $\boldsymbol{G}_S$ in (3.27); and the residual variance. Similarly, the SpATS model (for one time point), when genotypes are considered as fixed effects, has $(L+R+C+(b_2 b_3-1) = 180 + 60 + 28 + (8(8) - 1) =)$ 331 coefficients and 8 variance parameters (four less than in the previous case, due to the variances associated with the genotypes, one for each of the four populations). We note that when

genotypes are modelled as fixed effects, we decided not to introduce the population effect into the model to avoid identifiability problems. The computation time for each of both models (genotypes as random and fixed effects) was approximately 20 seconds.

As an output of the first stage, we obtained two inputs for the second stage of the two-stage approach: (i) the spatially corrected leaf area at the plant level (it included the estimated population and genotypic effects, as well as the residuals; see (4.4)); and (ii) the weights, which are used to propagate uncertainty from the first to the second stage (see (4.5)). Thus, to model the genetic signal in the second stage, we fitted the psHDM model (4.6) to the spatially corrected leaf area, with B-spline bases of dimension $b_{\text{pop}} = b_{\text{geno}} = b_{\text{plant}} = 11$ for population f_p, genotype f_g, and plant f_i functions. The mixed model formulation (4.13) of the psHDM has a total of 20152 regression coefficients (both fixed and random, $K \times b_{\text{pop}} + L \times b_{\text{gen}} + M \times b_{\text{plant}} = 4 \times 11 + 180 \times 11 + 1648 \times 11$) and 11 variance parameters (one for each population, four in total; three at genotype level – intercept, slope and smooth term – and the same at the plant level; plus the residual variance). Estimation took approximately 1 minute.

For the one-stage approach, we used the spatio-temporal psHDM (5.2) to fit the leaf area, with $b_{\text{row}} = b_{\text{col}} = 11$ B-splines for the random row, f_r, and column, f_c, effects; $b_1 = b_2 = b_3 = 8$ for the spatio-temporal smooth function, f_{ST}; and $b_{\text{pop}} = b_{\text{geno}} = b_{\text{plant}} = 11$ for the hierarchical components, f_p, f_g and f_i. Under this configuration, the mixed model formulation of the one-stage approach (see equation (5.12)) has a total of 21624 regression coefficients (both fixed and random, $K \times b_{\text{pop}} + L \times b_{\text{gen}} + M \times b_{\text{plant}} + R \times b_{\text{row}} + C \times b_{\text{col}} + (b_2 b_3 - 1)b_1 = 4 \times 11 + 180 \times 11 + 1648 \times 11 + 60 \times 11 + 28 \times 11 + ((8)(8) - 1) \times 8$) and 20 variance parameters (one for each population, four in total; three at genotype level – intercept, slope and smooth term – and the same at plant level, as well as for row and column effects; three variances in the row, column and time directions related with $f_{\text{ST}}(r, c, t)$; and the residual variance). Estimation took approximately 25 minutes.

7.1.2 PhenoArch results: One- and two-stage approaches comparison

We first start this section by raising the question about whether to model genotypes as fixed (as usually in stage-wise analyses) or random effects (as proposed in this thesis) when using the SpATS model (4.2) in the first stage of our two-stage approach (see Section 4.1.2). Figure 7.1 shows (for the plants of two genotypes, one per panel, under the two water regimes, as illustration) essentially identical results when we compare both results (for random genotypic effects curves are in blue and for fixed genotypic effects curves are in green) for the spatially corrected leaf area. We presume that this is due to the shrinkage of the genotypic BLUPs being counteracted by the inclusion of the residual component into the correction in equation (4.4). Henceforth, the results for the two-stage approach are based on genotypes modelled as random effects for its first stage.

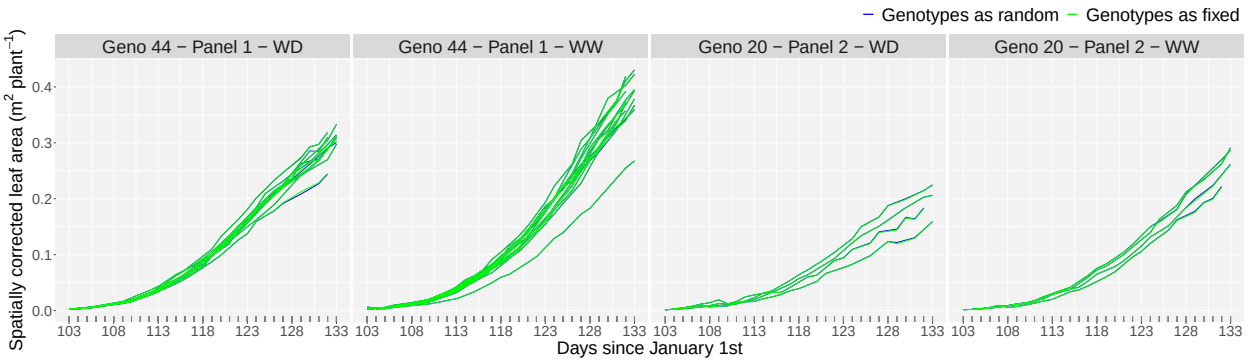


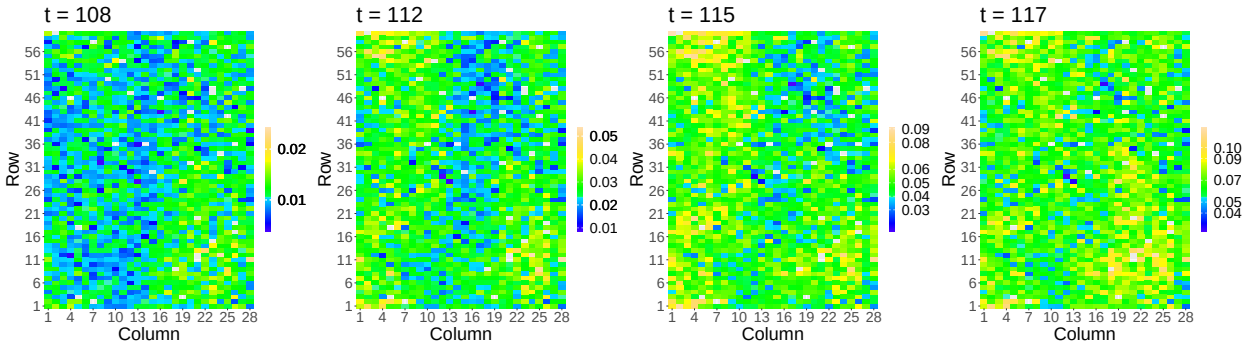
Figure 7.1: For the PhenoArch platform: Comparison of the evolution over time of the spatially corrected leaf area when modelling, in the first stage of the two-stage approach, genotypes as random (blue lines) or fixed (green lines) effects. Results are shown for the plants of two genotypes, one per panel, under the two water regimes (as illustration).

We follow the analysis by presenting the spatial trend results. As depicted in Figure 7.2, the spatial pattern observed in the raw data (see Figures 7.2(a) or 2.7) is successfully recovered by using both approaches (Figures 7.2(b) and (c)) at four different time points (as illustration). Figure 7.2(c) shows the centred spatial trends estimated by the one-stage approach (modelled through $f_{\text{ST}}(r, c, t)$ in (5.2)). As expected, they vary smoothly over time. In contrast, the spatial trends obtained with the two-stage approach (modelled in the first stage through $h_{\text{S}}(r, c)$ in (4.2) at time t) depicted in Figure 7.2(b) exhibit more marked differences among time measurements because analyses in the first-stage are performed separately per time point. Consequently, information on spatial heterogeneity is not shared across different measurement times. Finally, a detailed look at the scale of these two plots (estimated spatial trends) reveals small spatial effect for this particular dataset, when compared against the spatial distribution of the raw data in Figure 7.2(a).

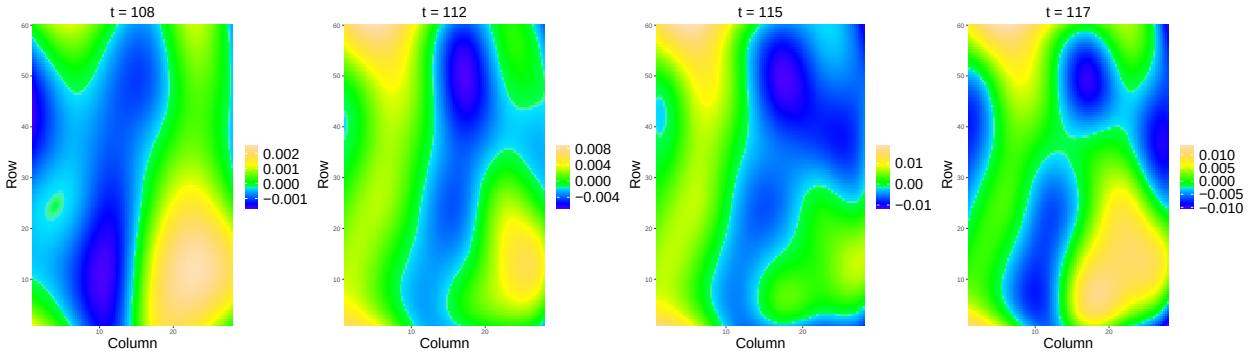
Once the spatial trend is modelled, results show that, as expected, the spatial pattern observed in the raw data (Figures 7.2(a) or 2.7) disappears in the spatially corrected phenotype, as illustrated in Figures 7.3(a) and (b) for the two- and one-stage approaches, respectively. Spatially corrected leaf area is obtained for the two-stage approach in the first stage through expression (4.4). Similarly, the spatially corrected leaf area for the one-stage approach can be calculated by eliminating from the spatio-temporal psHDM (5.2) the spatio-temporal trend, $f_{\text{ST}}(r, c, t)$, and the random row and column effects, i.e. $\tilde{y}_i(t) = \hat{f}_{p(i)}(t) + \hat{f}_{g(i)}(t) + \hat{f}_i(t) + \hat{\varepsilon}_i$.

Moreover, Figure 7.4 compares the evolution over time of the raw leaf area (grey lines) with the spatially corrected leaf area with the two-stage (blue lines) and one-stage (green lines) approaches. In general, the spatial correction reduced the variability among plants (i.e., replicates) of the same genotype and water treatment combination.

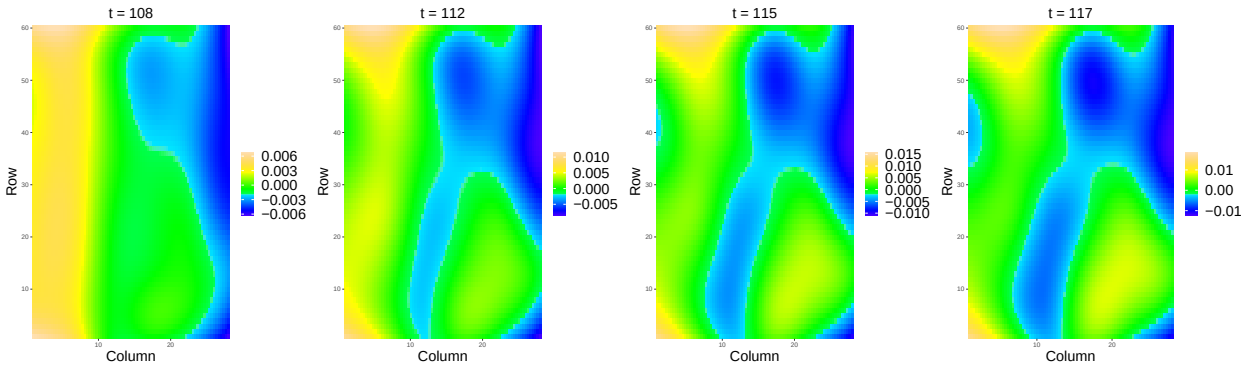
After the leaf area is spatially corrected, we can focus our analysis on the genetic signal. For both



(a) Raw spatial distribution

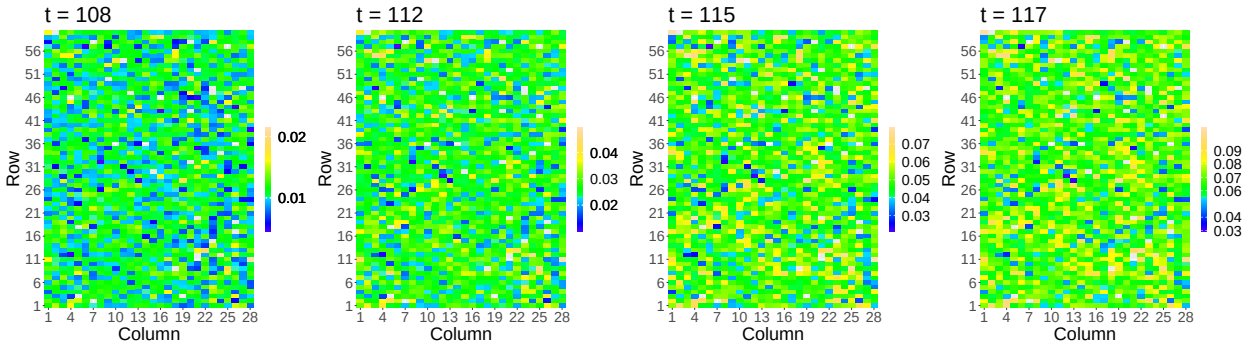


(b) Spatial trend two-stage approach

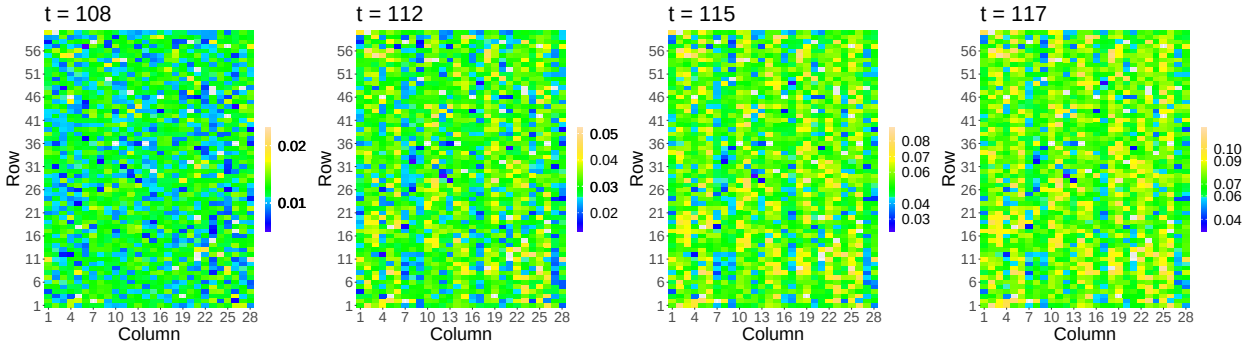


(c) Spatial trend one-stage approach

Figure 7.2: Results for the PhenoArch platform: (**a**) raw spatial distribution of the leaf area, and estimated spatial trend obtained with the (**b**) two-stage and (**c**) one-stage approaches, at four different measurements times ($t = 108, 112, 115, 117$ DOY). The colour scale is independently adjusted for each time point.



(a) Spatial distribution of the spatially corrected leaf area with the two-stage approach



(b) Spatial distribution of the spatially corrected leaf area with the one-stage approach

Figure 7.3: Results for the PhenoArch platform: Spatial distribution of the spatially corrected leaf area with the (**a**) two-stage approach and (**b**) one-stage approach at four different measurements times (t = 108, 112, 115, 117 DOY). The white areas denote missing data. The colour scale is independently adjusted for each time point.

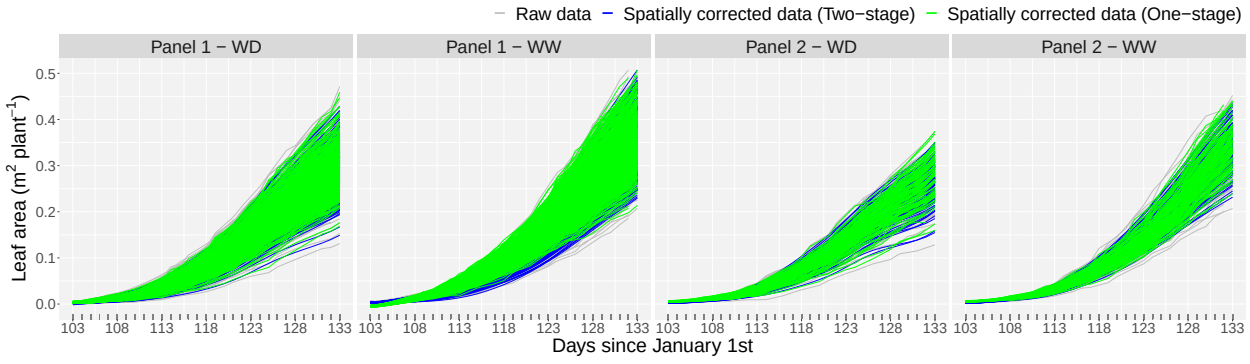


Figure 7.4: Results for the PhenoArch platform: Evolution over time of spatially corrected leaf area with the two-stage (blue lines) and one-stage (green lines) approaches. Results are depicted for all plants in the experiment as shown for the raw data (grey lines) in Figure 2.3.

approaches (for the two-stage approach from the psHDM (4.6) and for the one-stage approach from the spatio-temporal psHDM (5.2)) we analyse the following estimates at the three hierarchy levels

1. estimated population trajectories, $\hat{f}_p$, and respective first-order derivatives, $\hat{f}'_p$,
2. estimated genotype-specific deviations, $\hat{f}_g$, and respective trajectories, $\hat{f}_p + \hat{f}_g$, and first-order derivatives, $(\hat{f}_p + \hat{f}_g)'$, and
3. estimated plant-specific trajectories, $\hat{f}_p + \hat{f}_g + \hat{f}_i$.

Results for all genotypes are depicted in Figure 7.5: Figures 7.5(a) to (c) show results at population (orange and red lines for the one- and two-stage approaches, respectively) and genotype (with the one-stage approach in blue, and the two-stage approach in green) levels, and Figure 7.5(d) depicts results at genotype (continuous black lines for the one-stage and dotted black lines for the two-stage) and plant (grey continuous lines for the one-stage and grey dotted lines for the two-stage) levels for plants of two genotypes, one per panel, under the two water regimes (as illustration); blue lines correspond to raw data. Figure 7.5(a) shows a different growth pattern under both water regimes for the four populations (panel by water regime combination). That is, well watered (WW) plants grow faster than plants with water deficit (WD) for both panels. Consequently, the speed of leaf area growth (or leaf area growth rate, described by the estimated first-order derivative curves in Figure 7.5(b)) for WW plants reach higher values than those for WD plants. Genotype-specific deviations from their estimated population mean are shown in Figure 7.5(c)), where positive and negative deviations refer, respectively, to better and worse genotypic performance compared to the mean population. As expected, the magnitude of the deviations (and, thus, the differences in genotypes performance) increases with time. Also, genotypes from Panel 1 show the largest genetic variation under both water regimes. This is in concordance with the spatially corrected data showed in Figure 7.4. As illustrated with the estimated plant-specific trajectories in Figure 7.5(d), the two approaches are able to successfully recover the evolution over time of the raw leaf area in blue lines (recall that the grey lines represent spatially corrected data with both approaches, and then they can not be directly compared with the raw data, but they give an idea of the estimation accuracy), while appropriately handling the missing data. Moreover, genotype-specific trajectories (Figure 7.5(d), black lines) seem to summarise/describe the behaviour of the plant curves adequately. In the descriptive analysis in Section 2.1, we showed that missing data are mainly present in the second half of the observation period. We assume that the estimated curves in Figure 7.5 with the one-stage approach better describe the raw data structure in the presence of missing data when compare with estimated curves obtained with the two-stage approach. We believe this is due to the fact that the one-stage approach borrows strength across plant curves since less information is lost between and within stages. For instance, for this second half period, estimated trajectories with the two-stage approach are generally lower than curves with the one-stage approach (Figures 7.5 (a) and (c)). This period also shows the largest difference between the two approaches for the estimated first-order derivatives, which may affect the

precision with which growth and development related traits (e.g., local/global maximum/minimum speed values) can be extracted from the biomass trajectories. All in all, Figure 7.5 shows that the major discrepancy between the one- and two-stage approaches corresponds to the estimated first-order derivative curves in Figure 7.5(b). One- and two-stage approaches were most similar for genotypic deviations (Figure 7.5(c)). These results are consistent with the results of the simulation study (Section 6.3) and the analysis of other HTP datasets performed in the research period.

We now zoom in on the results at the genotype level, which is the decision-making level for plant breeders (genotypic performance can be assessed) and where the genotype-by-water regime interaction is analysed. In Figure 7.6, we use four genotypes per panel to illustrate our results for the one-stage (continuous lines) and two-stage (dotted lines) approaches. Genotypes were chosen such that two of them have the best (genotypes 48 and 15 in Panel 1 and 2, respectively) and worst (genotypes 27 and 20 in Panel 1 and 2, respectively) performance and the other two have an intermediate performance (genotypes 43 and 44 in Panel 1, and genotypes 03 and 29 in Panel 2). Results are essentially the same as the ones described for Figure 7.5. We additionally comment on the genotype-specific deviation results in Figure 7.6(c), which allows evaluating differences in genotypic performance. For instance, in Panel 1 and under both water regimes, genotype 48 (in pink) performs the best. This figure also allows us to analyse the genotype-by-water regime interaction. Note that genotypes 43 (in blue) and 44 (in purple) in Panel 1 have similar performance under WD, but their performances differ under WW. For Panel 2, we see that the curves for genotypes 03 (in red) and 15 (in brown) differ under WD, but become similar under WW. Differences between the one- and two-stage approaches for these curves are minimal.

7.1.3 PhenoArch results: Extracting time-independent attributes to characterise genotypes

One of the most important aspects in the previous analyses is whether decision-making changes with the approach used. To address this question, we extracted some time-independent features from the estimated curves to characterise genotypes (see Sections 4.2.7 and 5.5). We are aware that this data does not have information for the stationary phase common in the classic growth curve analysis. Consequently, the feature extraction is limited by the time window at which plants were measured. However, we calculate three features for all genotypes

1. maximum corrected leaf area (maxTrait) from the estimated genotype-specific trajectories (Figure 7.5(a)),

2. maximum speed rate (maxSpeed), before $t = 130$, from the estimated first-order derivatives for the genotype-specific trajectories (Figure 7.5(b)), and

3. area under the estimated genotype-specific deviations (AUC; Figure 7.5(c)).

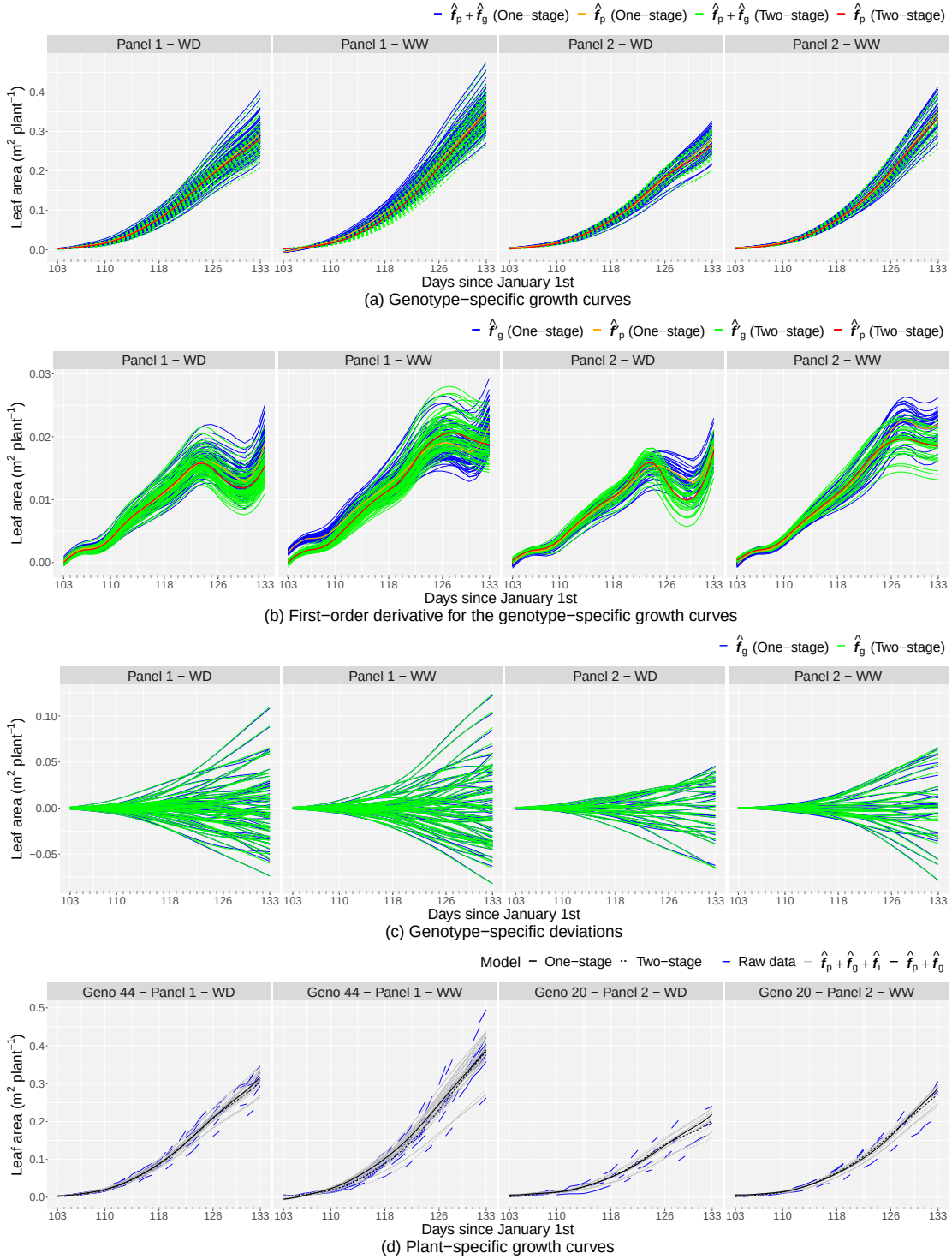


Figure 7.5: For the PhenoArch platform: For all genotypes, separately for each population and for both approaches: (**a**) estimated population- and genotype-specific trajectories, (**b**) estimated first-order derivative for the population- and genotype-specific trajectories, (**c**) estimated genotype-specific deviations, and (**d**) estimated genotype- and plant-specific trajectories (for plants of two genotypes, one per panel and under the two water regimes, as illustration).

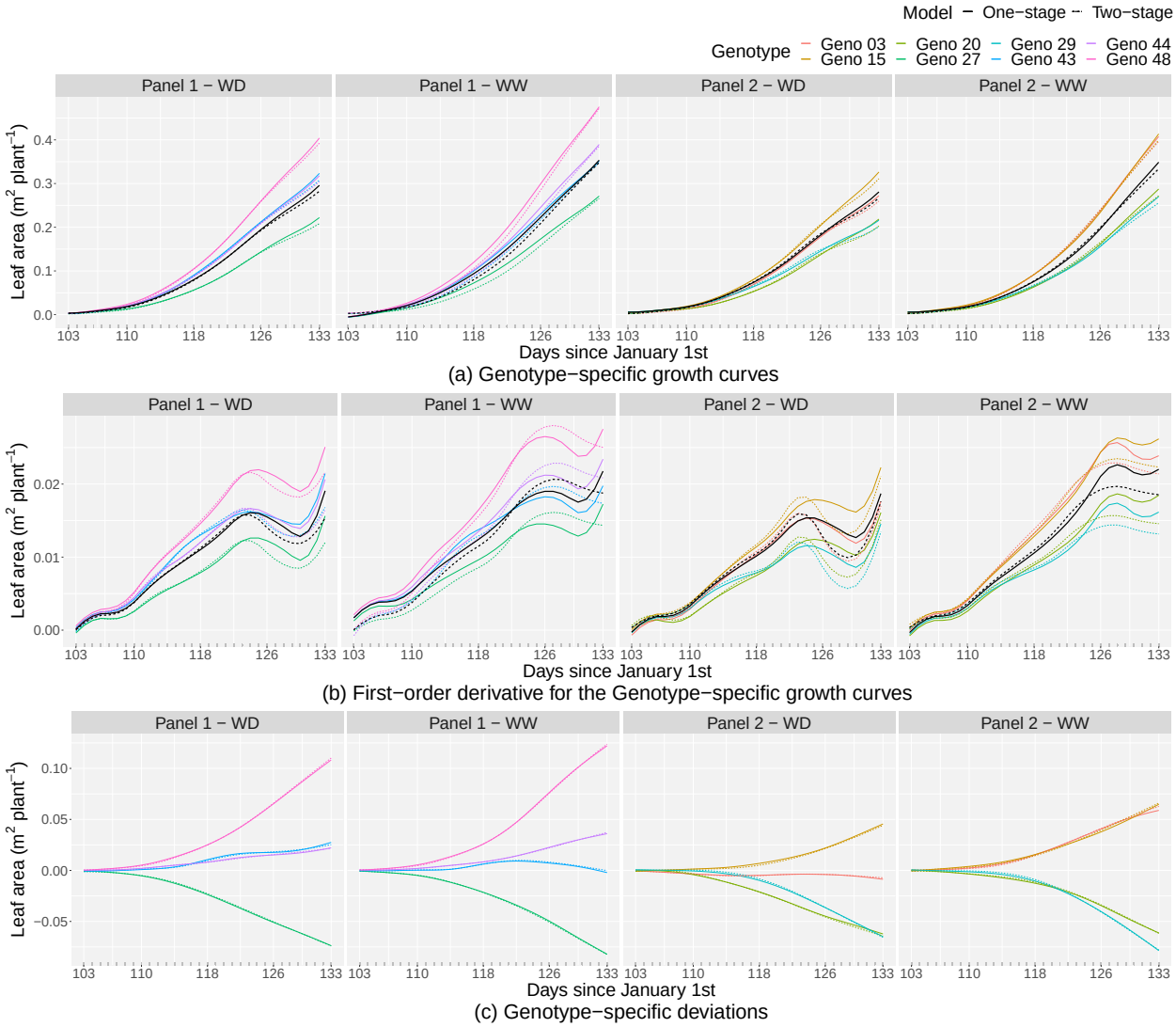


Figure 7.6: For the PhenoArch platform: Results at the genotype level (for four genotypes by panel, as illustration) for the one-stage (continuous lines) and two-stage (dotted lines) approaches, separately for each panel-by-water regime combination: (**a**) estimated genotype-specific trajectories, (**b**) estimated first-order derivatives for the genotype-specific trajectories, and (**c**) estimated genotype-specific deviations. In Figures (a) and (b) black lines represent curves at population level.

Bivariate scatter plots were used to compare the two approaches for each feature, as illustrated in Figure 7.7. The maxTrait and the AUC are the strongest correlated features, suggesting minimal differences in the decision-making process between both approaches. As expected and in concordance with the simulation results (Section 6.3), estimated first-order derivative curves are the most sensitive to differences between both approaches and consequently, the maxSpeed feature presents the largest differences between the one- and two-stage approaches (the results under the WW treatment show, for both panels, the highest difference).

However, the values of maxSpeed obtained using both approaches also show a high correlation within a panel-by-water regime combination. In brief, the results of both the simulation study and the application are consistent. That is, slight differences are detected when using the one- and two-stage approaches for most situations, with the estimation of the first-order derivative being the result with the most noticeable differences between the approaches proposed.

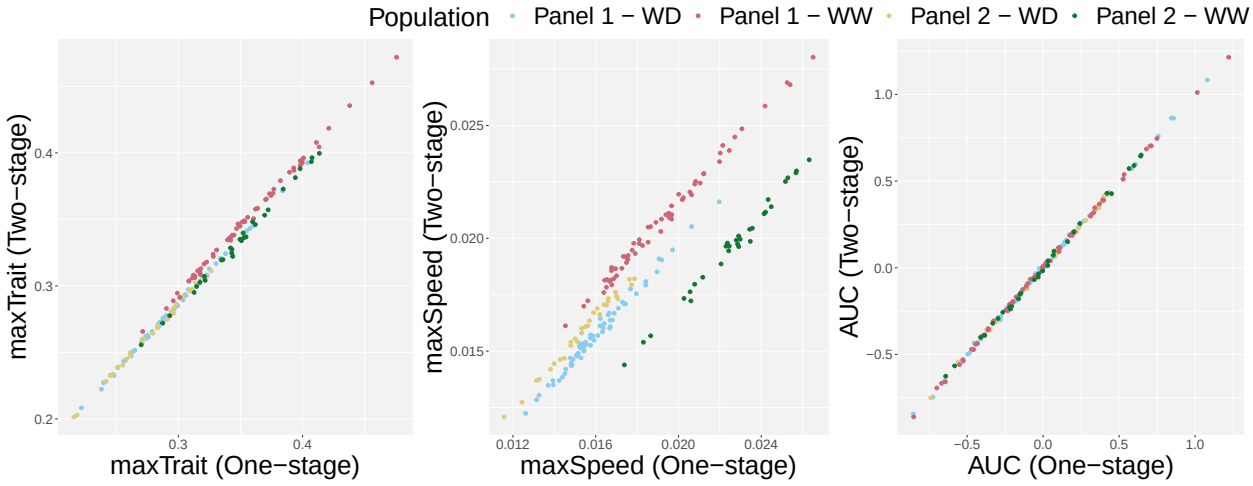


Figure 7.7: For the PhenoArch platform: Bivariate scatterplots with the extracted attributes at the genotype level. Each scatterplot depicts the comparison between the one- and the two-stage approaches for one feature. Colours represent the populations (panel-by-water regime combination).

7.2 FIP results

In Section 2.2, we described the data characteristics for the FIP platform. In contrast to the PhenoArch data, we have information available for this experiment for three trials (2015, 2016 and 2017, the experimental configuration of each trial was presented in Table 2.1). Thus, in addition to the genetic signal analysis, plant breeders are also interested in assessing the genotype consistency across trials (a total of 313 common genotypes) and identifying, e.g., those genotypes that perform the best.

7.2.1 FIP results: Approaches specification

We follow the same ideas presented before, for the PhenoArch data, to model the canopy height with the one- and two-stage approaches and for each trial separately. Once again, we comment on the number of coefficients and variance parameters to give the reader an idea of the complexity of our approaches. In this case, the SpATS model (4.1) used in the first stage of the two-stage approach included, besides the spatial trend $h_S(r, c)$, and the genotypic effects h_g (we later show differences in the fitted values when modelling h_g

as random (BLUPs) or fixed (BLUEs)), fixed effects for the two lots (experimental design factor h_e) and the seven wheat populations h_p (region of origin, with different genetic variance for each of the seven regions of origin when genotypes were considered as random effects), as well as random effects for the row and column positions, h_r and h_c. For the spatial trend $h_S(r, c)$, $b_2 = b_3 = 8$ B-spline basis dimensions were assumed for the row and column positions of the virtual grid, respectively. The correction (see equation (4.4)) included the estimated population (when genotypes were random), the genotypic effects and the residuals, and we averaged over the lot fixed effect to eliminate its impact.

Consequently, when modelling genotypes as random, the mixed model formulation (4.2) of the SpATS model has 475, 472 and 489 coefficients and 15 variance parameters (seven genotype variances associated with each region, two variances for the random row and column effects, five variances for the smooth spatial function, and the residual variance) for each trial (2015, 2016 and 2017) at each time point, respectively. Similarly, when genotypes were considered as fixed effects, the SpATS model, has 434, 431, and 448 coefficients and 8 variance parameters for each trial (2015, 2016 and 2017) at each time point, respectively. Note that, for each time point we have seven variance parameters less than in the previous case, due to the variances associated with the genotypes, one for each of the seven regions of origin. The computation time for each of both models (genotypes as random and fixed effects) and each of the three trials was approximately 20 seconds.

In the second stage of the two-stage approach, we used the spatially corrected canopy height to model the genetic signal. The trajectories for the spatially corrected phenotype show here a more complex pattern than for the PhenoArch platform (see Figures 2.3 and 2.11), so we used the psHDM (4.6) with cubic B-spline bases of dimension $b_{\text{pop}} = b_{\text{geno}} = b_{\text{plant}} = 20$ for the three levels of the hierarchy (f_p, f_g and f_i). The mixed model formulation (4.13) of the psHDM has a total of 20180, 20080, and 21220 regression coefficients (both fixed and random) and 14 variance parameters (one for each of the seven regions of origin, three at genotype level – intercept, slope and smooth term – and the same at the plant level, and the residual variance) for each trial (2015, 2016 and 2017), respectively. The fitting processes needed approximately 1 minute for each of the three trials.

Following the modelling choices made for the two-stage approach, for the one-stage, we used the spatio-temporal psHDM (5.2) to fit the canopy height, with $b_{\text{row}} = b_{\text{col}} = 20$ B-splines for the random row, f_r, and column, f_c, effects, $b_1 = b_2 = b_3 = 8$ for the spatio-temporal smooth function, f_{ST}, and $b_{\text{pop}} = b_{\text{geno}} = b_{\text{plant}} = 20$ for the hierarchical components, f_p, f_g and f_i. Under this configuration, the mixed model formulation of the one-stage approach (see the spatio-temporal psHDM (5.2)) has a total of 22204, 22104 and 23284 regression coefficients (both fixed and random) and 20 variance parameters (one for each of the seven regions of origin; three at genotype level – intercept, slope and smooth term – and the same at plant level, as well as for row and column effects; three variances in the row, column and time directions related with $f_{\text{ST}}(r, c, t)$; and the residual variance) for each trial (2015, 2016 and 2017), respectively. Although we

could explore a more general one-stage approach, our current formulation of the spatio-temporal psHDM (5.2) and its implementation (code) is very specific. We only consider the spatio-temporal smooth function, f_{ST}, and random effects for rows and columns, f_r, and f_c, as non-genetic effects, but no other experimental factors are taken into account. Thus, in contrast to the two-stage approach, the lot effect is not included into the one-stage approach. Estimation took approximately 35 minutes for the 2015 trial, 1 hour and 30 minutes for the 2016 trial (it took more iterations until convergence), and 30 minutes for the 2017 trial.

7.2.2 FIP results: One- and two-stage approaches comparison

We start by comparing in Figure 7.8 the spatially corrected canopy height obtained when modelling genotypes either as fixed (green lines) or random (blue lines) effects for the plants of one genotype by region of origin (as illustration) and for the three trials. As for the PhenoArch data (see Figure 7.1), the results are very similar. In the following results, genotypes are modelled as random effects for the first stage of the two-stage approach.

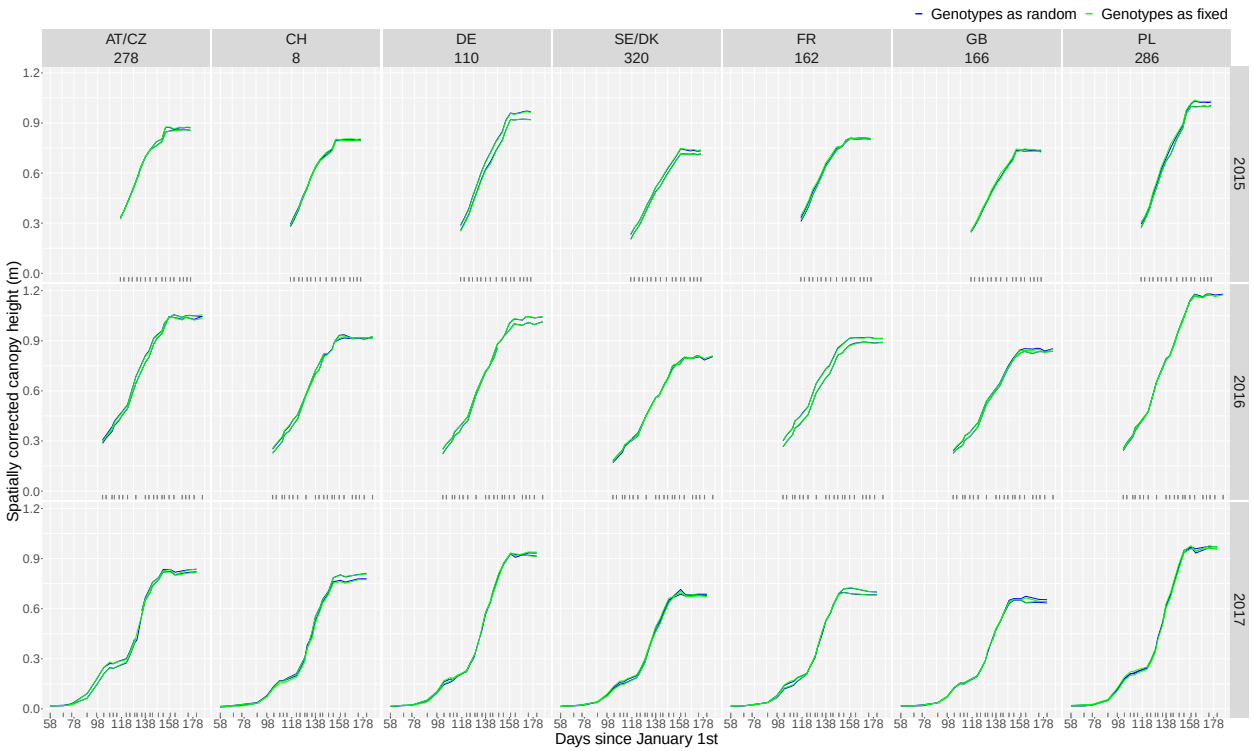
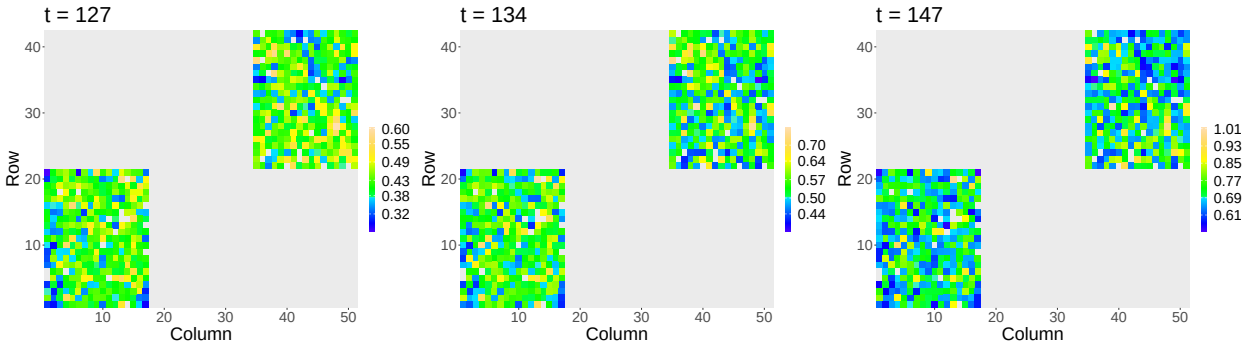


Figure 7.8: For the FIP platform: Comparison of the evolution over time of the spatially corrected canopy height when modelling, in the first stage of the two-stage approach, genotypes as random (blue lines) or fixed (green lines) effects. Results are shown for the plants of one genotype per region of origin (as illustration). AT/CZ: Austria/Czechia; CH: Switzerland; DE: Germany; FR: France; GB: Great Britain; PL: Poland; SE/DK: Sweden/Denmark.

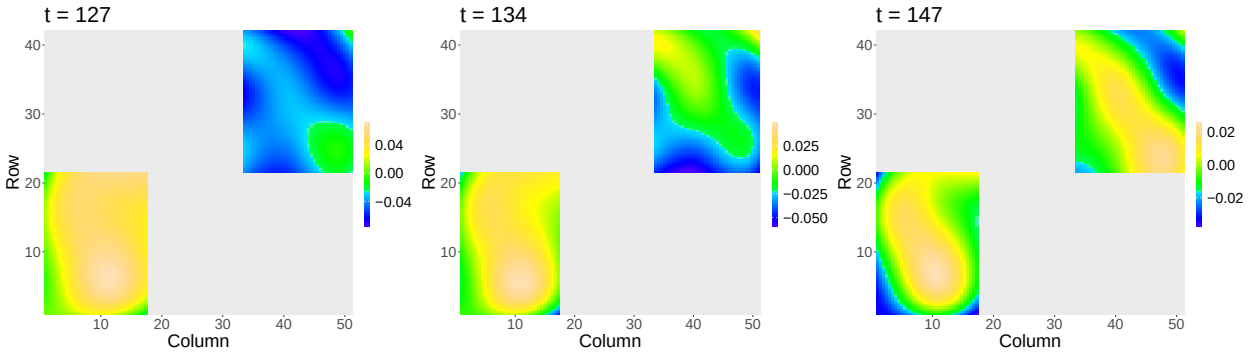
Figures 7.9, 7.10 and 7.11 depict the spatial trend obtained when modelling the canopy height for the three trials (2015, 2016 and 2017, respectively) with the one- and two-stage approaches at three different measurement days (DOY). As for the PhenoArch data, the spatial trends vary smoothly over time when the one-stage approach is used (Figures 7.9(c), 7.10(c) and 7.11(c)), and for the two-stage approach (Figures 7.9(b), 7.10(b) and 7.11(b)) differences in the spatial trend are more evident through time. A look at the canopy height scale for these plots shows a small spatial effect. We note that a particularity of this experiment is that the two lots for the three trials are non-adjacent. In this case, we fill in the gaps with NA (where necessary) to assume the complete grid.

As for the PhenoArch experiment, the correction performed in the first stage reduced the variability among replicates of the same genotype. In Figure 7.15, we show all the replicates for each region of origin by trial. We observe that the variability among raw canopy height (green lines) plants of the same region of origin and trial are higher than for the spatially corrected data with the one-stage (blue lines) and the two-stage approaches (in grey). This reduction is due to the lot effect and the spatial variation when using the two-stage approach and only due to the spatial variation when using the one-stage approach

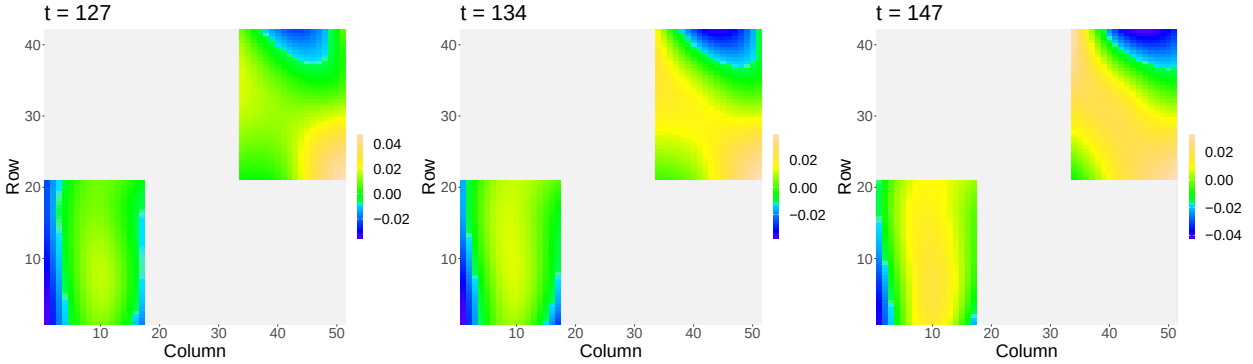
Once the spatially corrected canopy height is obtained, we analyse the genetic signal with the two proposed approaches. To that aim, we show for each region of origin and trial the estimated region- and genotype-specific trajectories in Figure 7.16, their estimated first-order derivatives in Figure 7.17, and the estimated genotype-specific deviations in Figure 7.18. Results show differences between regions of origin (and between genotypes within a region) in, e.g., growth patterns (Figure 7.16), growth rates (Figure 7.17), and genotype performance (Figure 7.18) for the three trials. In contrast to the PhenoArch data, for the three trials of this experiment, we found more than one maximum point in their first-order derivatives (Figure 7.17). For instance, for the 2017 trial, these speed rates correspond to maxima around DOYs 97, 133 and 147, respectively. We have information about the mean temperature for this trial, as shown in Figure 2.15 (blue line). Thus, maximum speed rates around DOY 133 and 189 can be interpreted, respectively, as recoveries after a severe cold period in April (DOY 110-120) and a milder one in May (DOY 140). However, for DOY 189 the growth rates declined as plants approached their final height. We also observe a wigglier behaviour (but more similar through regions of origin) of the first-order derivatives for the 2016 trial than for the 2015 and 2017 trials. Growth rates reach higher values for Austria/Czechia (AT/CZ), Switzerland (CH), Germany (DE) and Poland (PL) than for Sweden/Denmark (SE/DK), France (FR) and Great Britain (GB) for both, 2015 and 2017, trials. We also highlight the importance of the genotype-specific deviations (Figure 7.18) when comparing the performance among genotypes of the same region. Deviations refer to the "pure" genetic signal since the non-genetic effects (spatial effects and other experimental factors) and the regional trends are removed from the phenotype of interest. We observe that genetic effects are minimal at the beginning of the experiment, but they become higher over time. Moreover, we observe some regions of origin with late deviations (e.g. genotypes in the Austria/Czechia (AT/CZ) region for the 2015 and 2016 trials) and others with early deviations (e.g. genotypes in the Germany (DE) region for the 2015 and 2016



(a) Raw spatial distribution



(b) Spatial trend two-stage approach

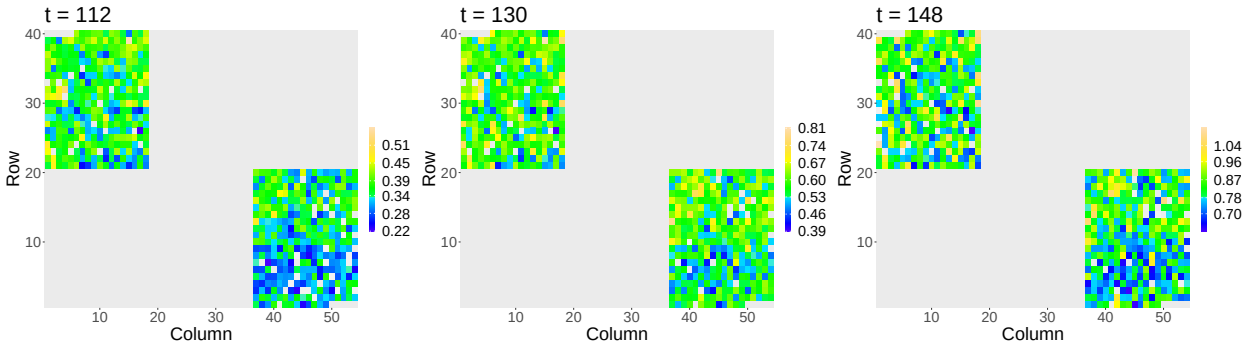


(c) Spatial trend one-stage approach

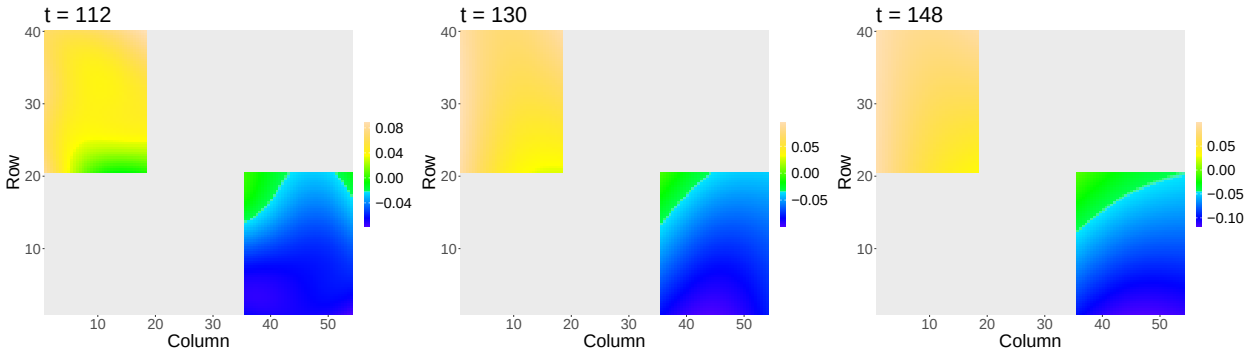
Figure 7.9: Results for the FIP platform, 2015 trial: (**a**) raw spatial distribution of the canopy height, and estimated spatial trend obtained with the (**b**) two-stage and (**c**) one-stage approaches ($t = 127, 134, 147$ DOY). The colour scale is independently adjusted for each time point.

trials). Thus, the AUC (see Section 4.2.7) becomes a good indicator of genotype performance.

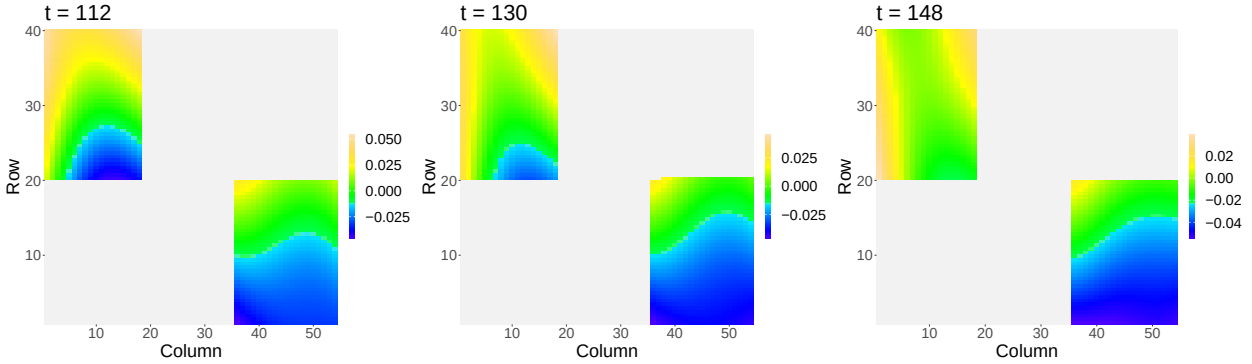
We finish this section by commenting on the comparison between the two approaches. We found small differences between the one- and two-stage approaches for the three kinds of curves. Following the results of the simulations and the PhenoArch data, the estimated first-order derivatives are the curves with the



(a) Raw spatial distribution



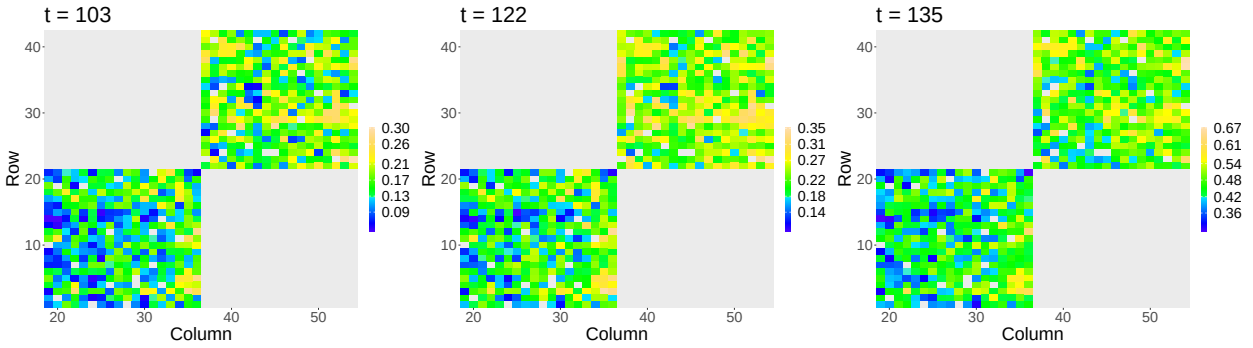
(b) Spatial trend two-stage approach



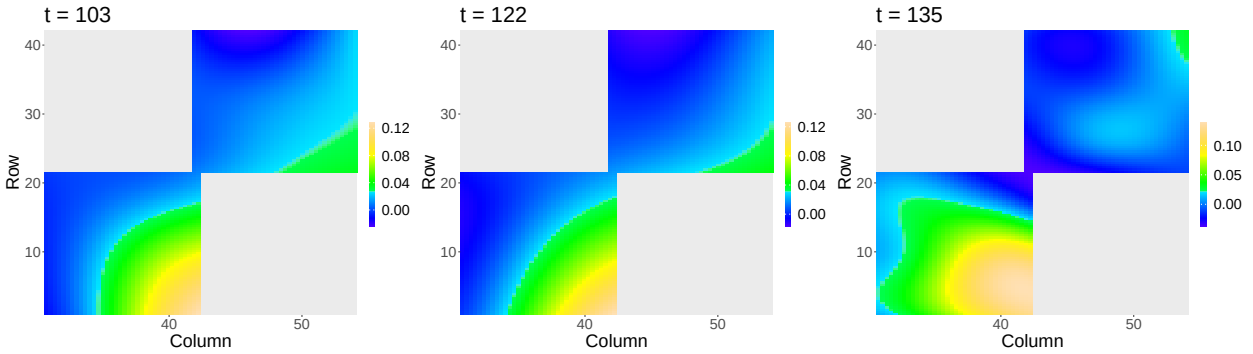
(c) Spatial trend one-stage approach

Figure 7.10: Results for the FIP platform, 2016 trial: (**a**) raw spatial distribution of the canopy height, and estimated spatial trend obtained with the (**b**) two-stage and (**c**) one-stage approaches ($t = 112, 130, 148$ DOY). The colour scale is independently adjusted for each time point.

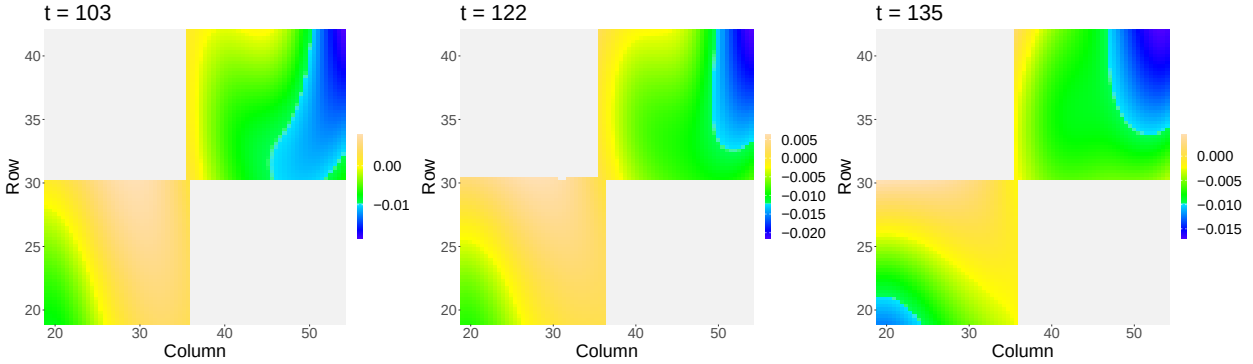
biggest differences (with special attention on the 2016 trial). A large amount of missing data characterised phenoArch data. In contrast, time series curves at irregularly spaced time points characterise FIP data. In general, the 2015 trial shows more consistent results between the two approaches than the 2016 and 2017 trials, which have more spaced measurement times. We believe that the one-stage approach outperforms



(a) Raw spatial distribution



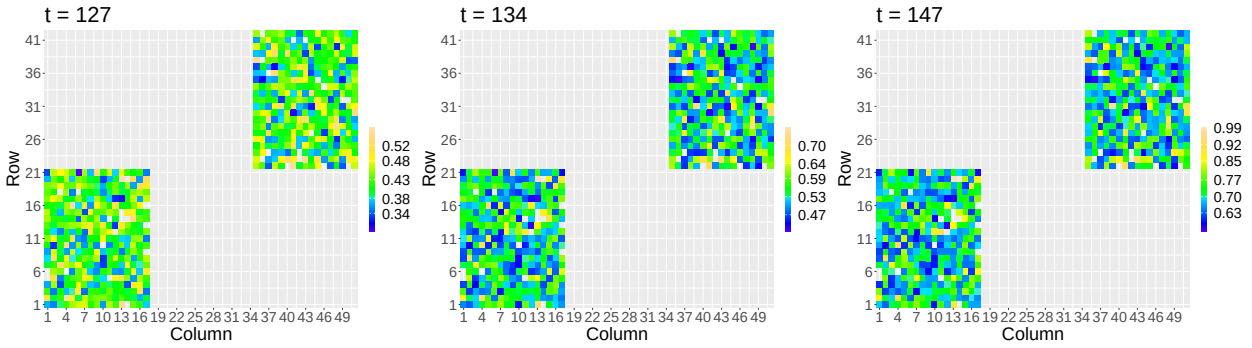
(b) Spatial trend two-stage approach



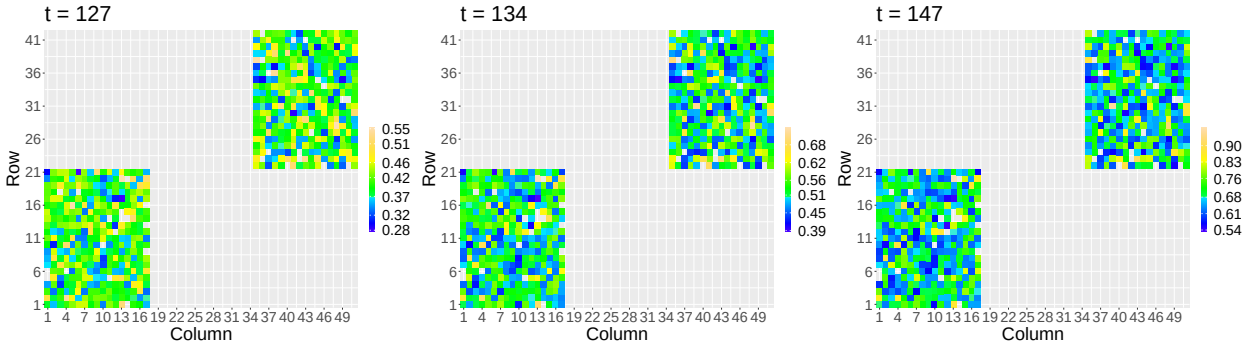
(c) Spatial trend one-stage approach

Figure 7.11: Results for the FIP platform, 2017 trial: (**a**) raw spatial distribution of the canopy height, and estimated spatial trend obtained with the (**b**) two-stage and (**c**) one-stage approaches ($t = 103, 122, 135$ DOY). The colour scale is independent for each time point.

the two-stage approach both in the presence of missing data and irregularly spaced time points. However, further exploration in this direction is required (e.g., through simulated data).



(a) Two-stage approach



(b) One-stage approach

Figure 7.12: Results for the FIP platform, 2015 trial: Spatial distribution of the spatially corrected canopy height with the (**a**) two-stage and (**b**) one-stage approaches (t = 127, 134, 147 DOY). The white areas denote missing data. The colour scale is independent for each time point.

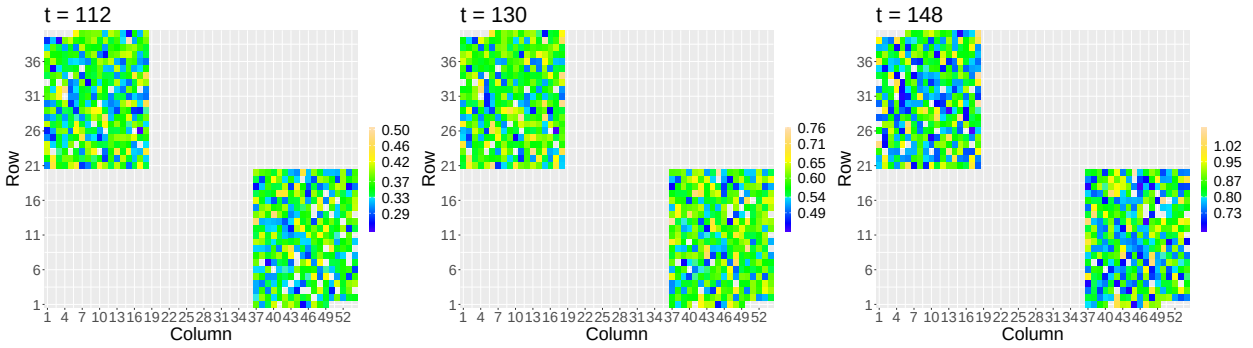
7.2.3 FIP results: Extracting time-independent attributes to characterise genotypes

In addition to characterising the genotypes, in this section, we aim to assess the genotype consistency across trials. For that purpose, we extracted, for the 313 common genotypes to the three trials, the same three features that we used for the PhenoArch analysis in Section 7.1.3, that is

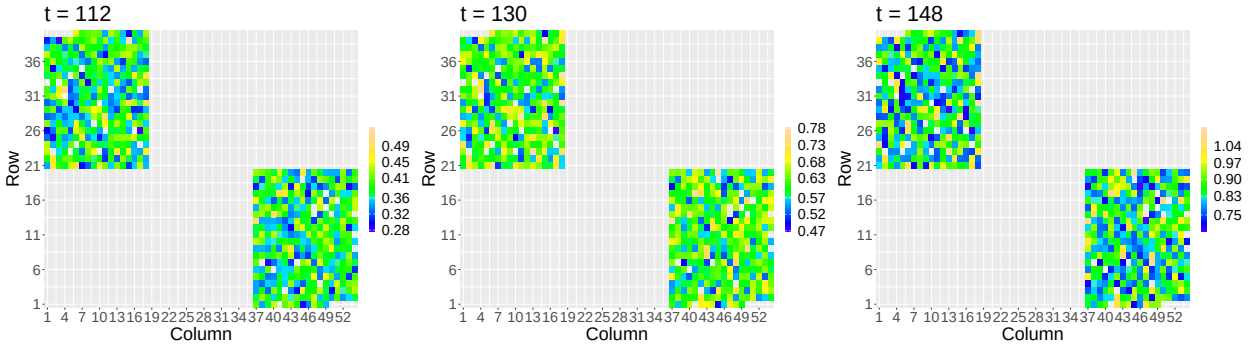
1. maximum corrected canopy height (maxTrait) from the estimated genotype-specific trajectories (Figure 7.16),

2. maximum speed rate (maxSpeed) between $118 \leq t \leq 148$ from the estimated first-order derivatives for the genotype-specific trajectories (Figure 7.17). We are aware that more than one maxima point can be obtained for the three trials, but for the sake of simplicity, we extract only one for the time window common to the three trials in which the plants grew the most, and

3. area under the genotype-specific deviations (AUC; from Figure 7.18). For a fair comparison between



(a) Two-stage approach

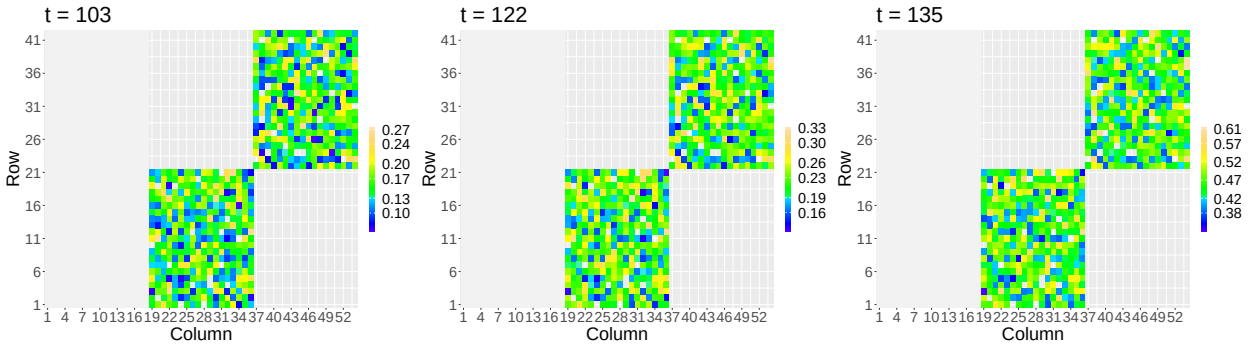


(b) One-stage approach

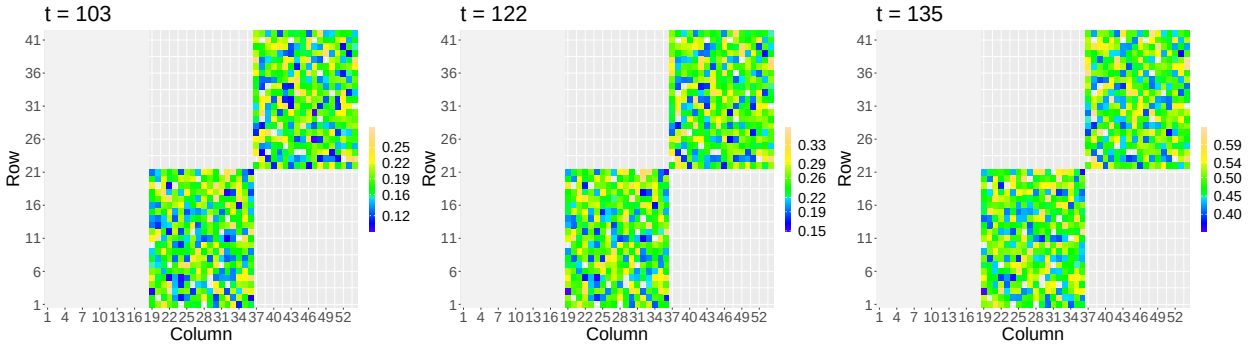
Figure 7.13: Results for the FIP platform, 2016 trial: Spatial distribution of the spatially corrected canopy height with the (**a**) two-stage and (**b**) one-stage approaches (t = 112, 130, 148 DOY). The white areas denote missing data. The colour scale is independent for each time point.

trials, the AUC was calculated for the time interval $117 \leq t \leq 176$, which is the common time window where the genotypes were measured for the three trials (see Table 2.1). Nothing, however, precludes focusing attention on a different time interval.

The two approaches are compared by bivariate scatter plots for each feature and trial, as shown in Figure 7.19. The bivariate scatterplots of the extracted genotype-specific attributes show that the genotypes cluster according to their region of origin. Results for this platform are consistent with those previously presented for the PhenoArch platform (Section 7.1): strongly correlated features, indicating small differences between both approaches for the decision-making process. The results for the maxSpeed for the 2016 trial show the highest difference (but with a high correlation between both approaches within a region of origin). For each attribute, region of origin and trial, we identified the genotype with the maximum value (i.e., the "best" genotype) for the two-stage (genotype numbers to the left and in bold) and the one-stage (genotype numbers to the right) approaches. We found that in 9 of the 63 (= 3 attributes $\times$ 3 trials $\times$ 7 regions of origin) comparisons, the genotype with the best performance is not the same for the two approaches (with the Great



(a) Two-stage approach



(b) One-stage approach

Figure 7.14: Results for the FIP platform, 2017 trial: Spatial distribution of the spatially corrected canopy height with the (**a**) two-stage and (**b**) one-stage approaches (t = 103, 122, 135 DOY). The white areas denote missing data. The colour scale is independent for each time point.

Britain (GB) region of origin and the maxSpeed attribute the least consistent). The most consistent region of origin is Switzerland (CH): the same genotype (genotype 20) is identified by the two approaches, and the same genotype has the best performance for the three attributes (except for the maxTrait in the 2016 trial).

Univariate analysis of each attribute using boxplots to compare regions of origin within trials with each approach is shown in Figure 7.20 (for the 313 common genotypes to the three trials). These boxplots also show regional clustering. For instance, for the maxTrait, three clusters of regions of origin are identified (they are consistent through trials and for the two approaches): fast-growing genotypes (Austria/Czechia (AT/CZ) and Poland(PL), eastern regions), medium-growing genotypes (Switzerland (CH) and Germany (DE), central regions), and low-growing genotypes (Franc (FR), Great Britain (GB) and Sweden/Denmark (SE/DK), western and northern regions). Small changes are observed between the one- and two-stage approaches for these comparisons, being the maxSpeed for the 2016 trial with the most different results.

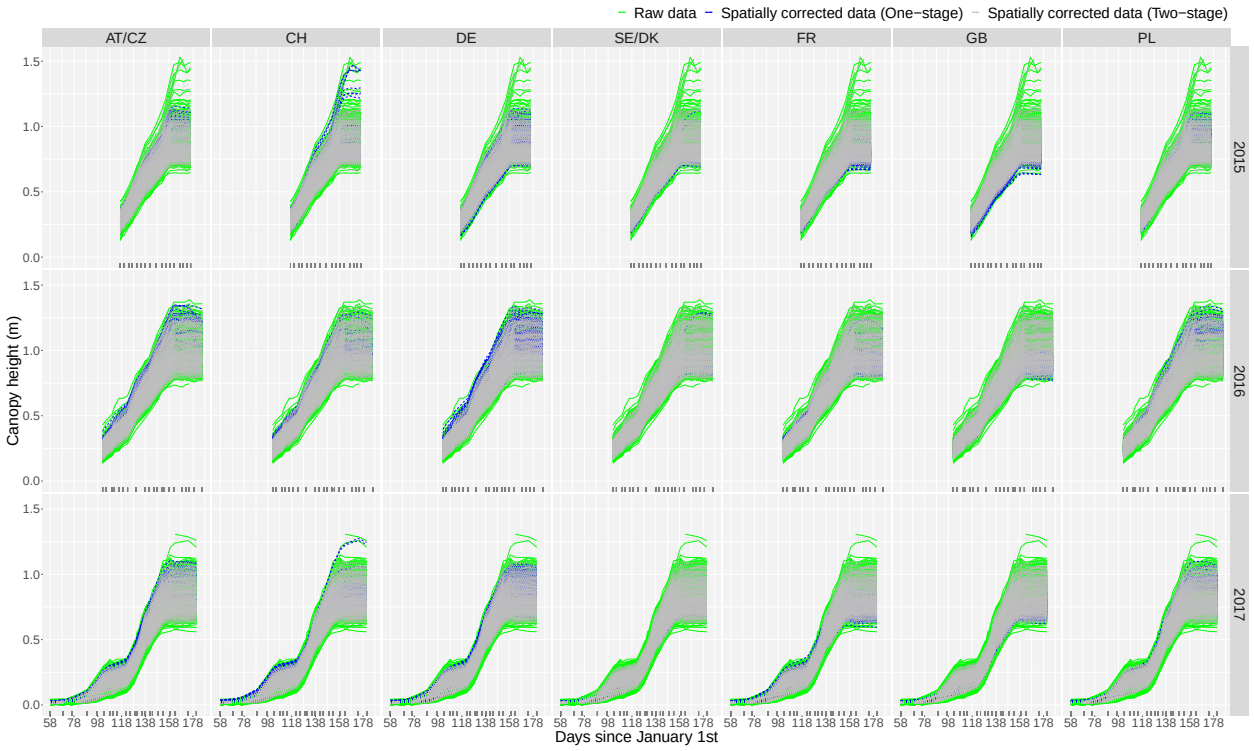


Figure 7.15: Results for the FIP platform: Evolution over time of spatially corrected canopy height with the two-stage (grey lines) and one-stage (blue lines) approaches. Results are shown for all plants in the experiment as shown for the raw data (green lines) in Figure 2.11. AT/CZ: Austria/Czechia; CH: Switzerland; DE: Germany; FR: France; GB: Great Britain; PL: Poland; SE/DK: Sweden/Denmark.

7.2.4 FIP results: Use of time-independent attributes to characterise regional adaptation

While a deeper physiological analysis is beyond the scope of this thesis, we will use the extracted attributes to highlight the potential benefit of an in-depth analysis of spline-based growth patterns. Here, we use the regional groups shown in Figure 7.21, but a similar analysis could be done using individual genotypes. For instance, for the 2017 trial, the observed height development follows a principally logistic growth curve: stem elongation started after the plants were vernalised over winter (by means of cold exposure) and ended around flowering. However, the height development plateaued between days 103 and 118 in all genotypes, most likely due to a cold period in April (see available information for the mean temperature (blue line) for this trial in Figure 2.15). When looking at the region-specific trajectories, this short and extreme phase caused even rank changes in growth ($\hat{f}_p$ in Figure 7.21): the regional groups showing most vigorous growth before the stress (first local maxima point in $\hat{f}'_p$ in Figure 7.21) stopped growth completely while those which grew slowest could maintain some growth during the cold (first local minima following the first local maxima). Such pattern may point to physiological adaptations to the different climatic regions of Europe as the slow-growing northern types from Great Britain (GB), and Denmark and Sweden (SE/DK) showed

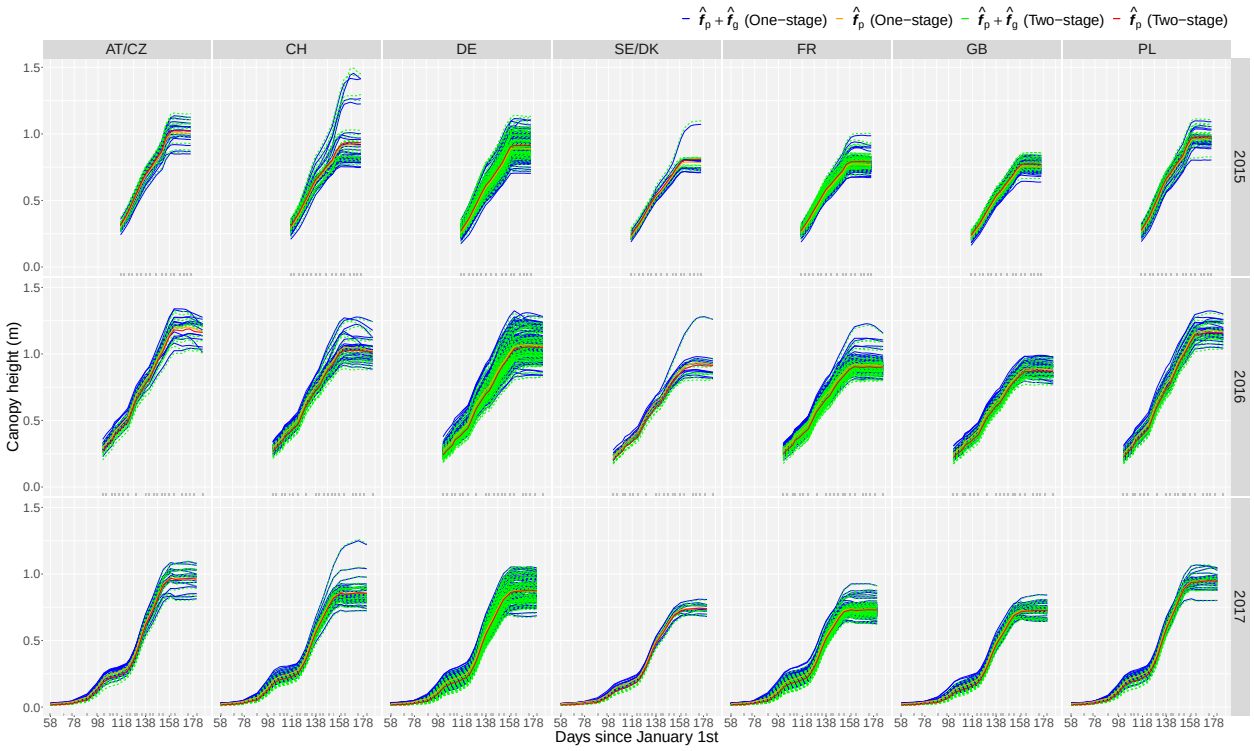


Figure 7.16: For the FIP platform: Estimated region- and genotype-specific trajectories for all genotypes, separately for each region of origin, trial and for both approaches. AT/CZ: Austria/Czechia; CH: Switzerland; DE: Germany; FR: France; GB: Great Britain; PL: Poland; SE/DK: Sweden/Denmark.

least response to cold while the fast-growing continental types from Poland (PL) and Austria and Czechia (AT/CZ) stopped growing. Moreover, the genotypes from the southwest – France (FR) and Switzerland (CH) – did not recover growth up to the same level as the more northern and eastern varieties did (compare first and second local maxima point). A multi-year analysis shows that the region-specific average development is consistent through trials, with changes through time. We highlight, e.g., the performance of the Poland (PL) region, which consistently (for the three trials) started with moderate growth and finished with one of the most higher recovery growth rates. Regional comparison of the results between the one- and two-stage approaches shows slight differences.

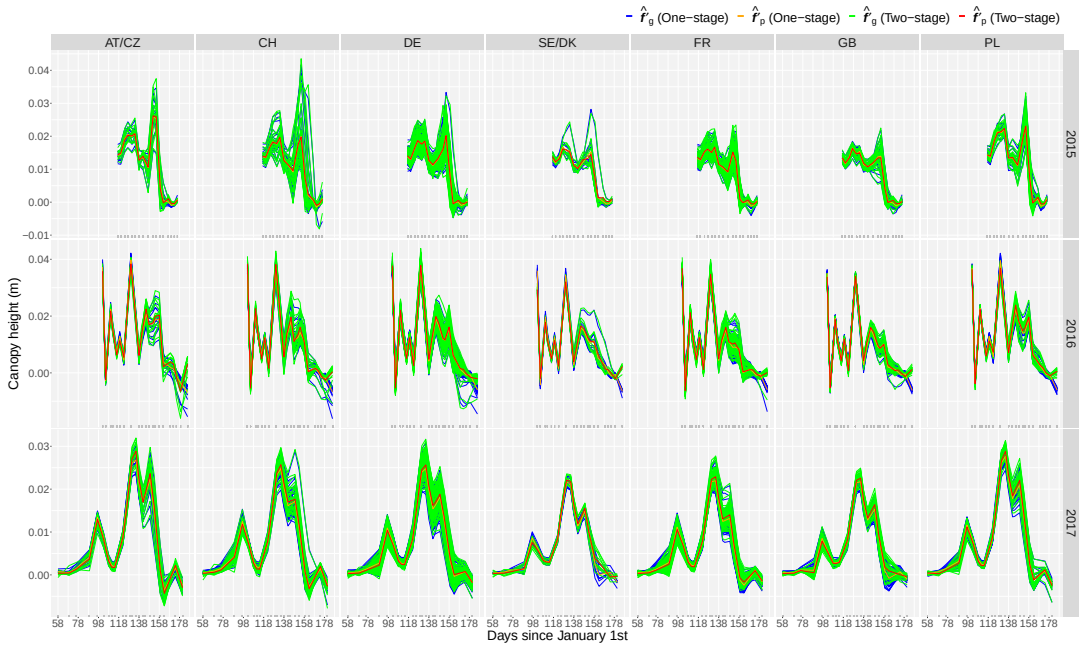


Figure 7.17: For the FIP platform: Estimated region- and genotype-specific first-order derivatives for all genotypes, separately for each region of origin, trial and for both approaches. AT/CZ: Austria/Czechia; CH: Switzerland; DE: Germany; FR: France; GB: Great Britain; PL: Poland; SE/DK: Sweden/Denmark.

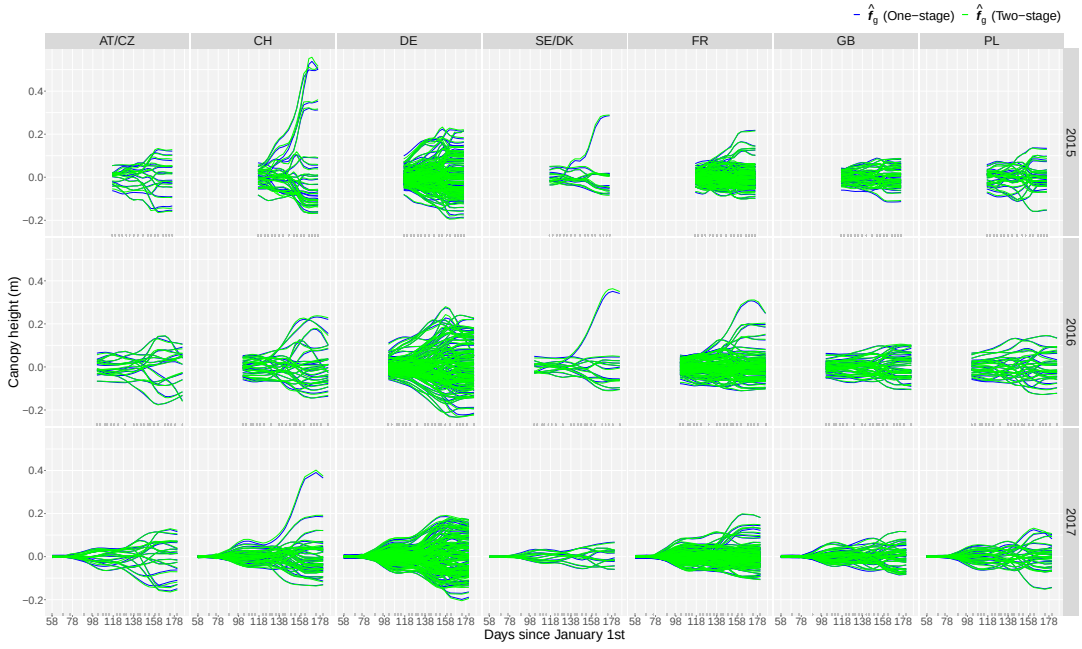


Figure 7.18: For the FIP platform: Estimated genotype-specific deviations for all genotypes, separately for each region of origin, trial and for both approaches. AT/CZ: Austria/Czechia; CH: Switzerland; DE: Germany; FR: France; GB: Great Britain; PL: Poland; SE/DK: Sweden/Denmark.

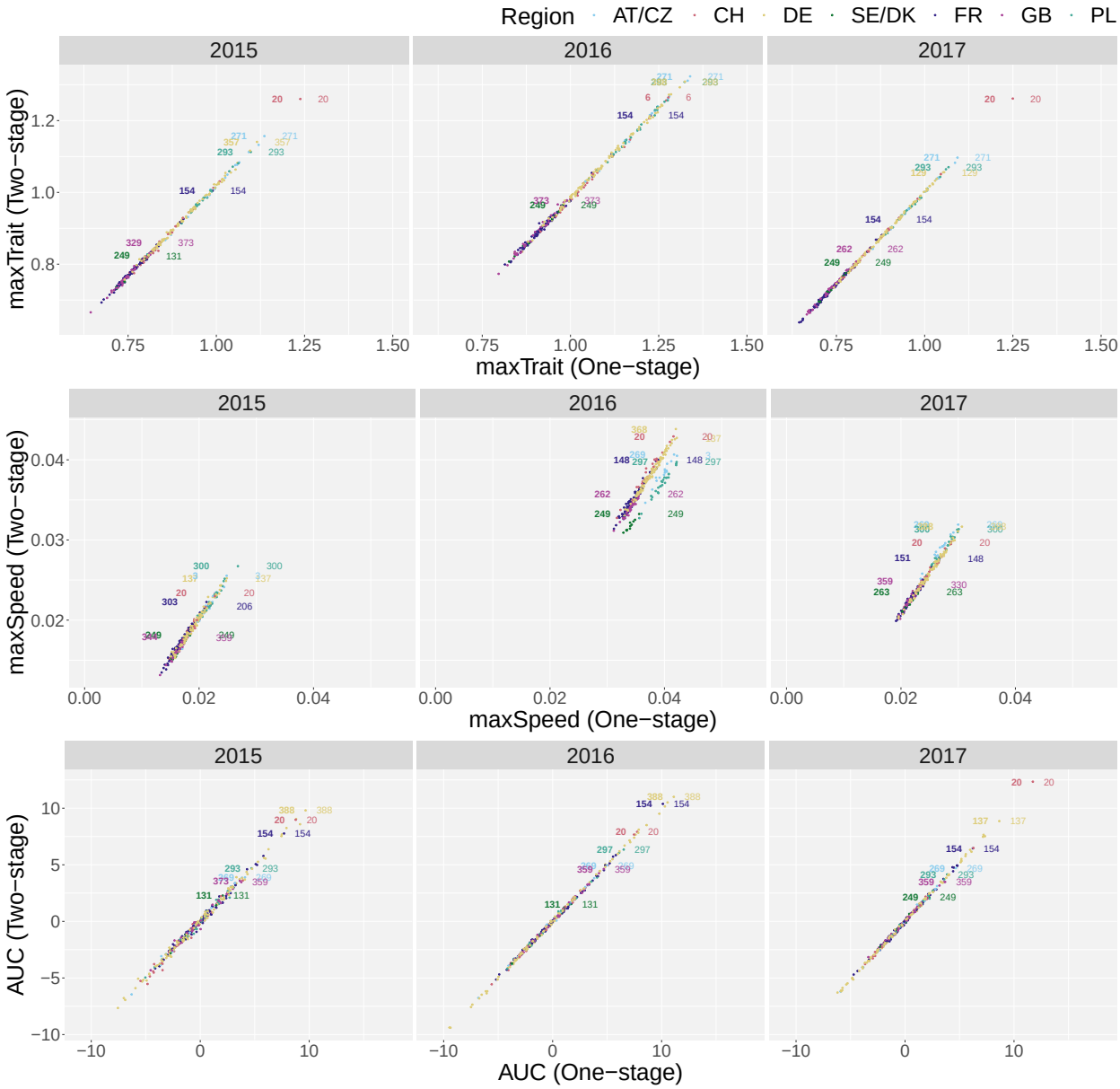


Figure 7.19: For the FIP platform: For the 313 common genotypes for the three trials, bivariate scatter plots with the extracted attributes at the genotype level. Each scatterplot depicts the comparison between the one- and the two-stage approaches for one feature in one trial. Colours represent the regions of origin. Points identified with text are the genotypes with the maximum feature value ("best" genotypes, one by region of origin). Genotypes to the left and in bold stands for the two-stage approach. AT/CZ: Austria/Czechia; CH: Switzerland; DE: Germany; FR: France; GB: Great Britain; PL: Poland; SE/DK: Sweden/Denmark.

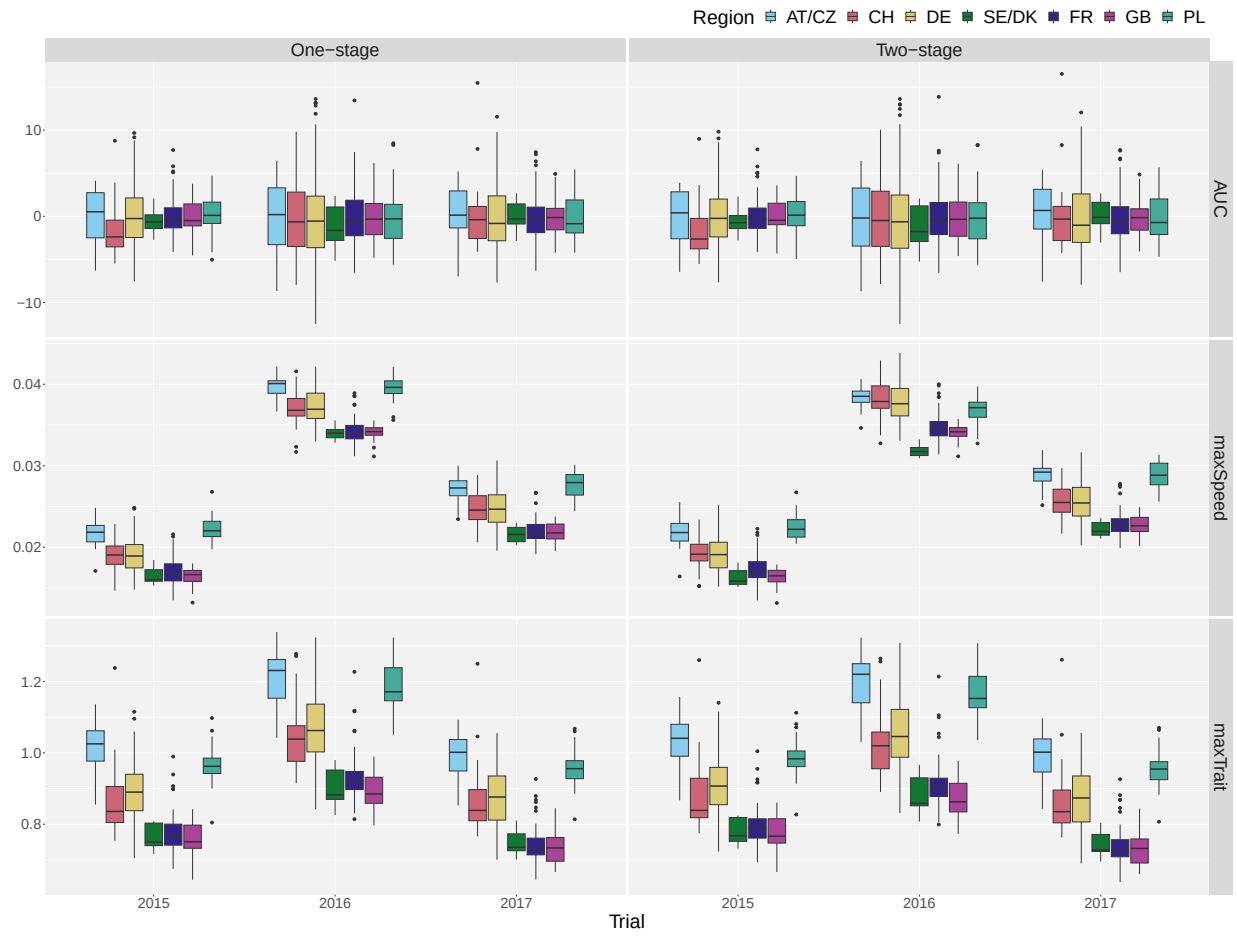


Figure 7.20: For the FIP platform: For the 313 common genotypes for the three trials, boxplots with the extracted attributes at the genotype level by trial, region of origin and approach. Colours represent the regions of origin. AT/CZ: Austria/Czechia; CH: Switzerland; DE: Germany; FR: France; GB: Great Britain; PL: Poland; SE/DK: Sweden/Denmark.

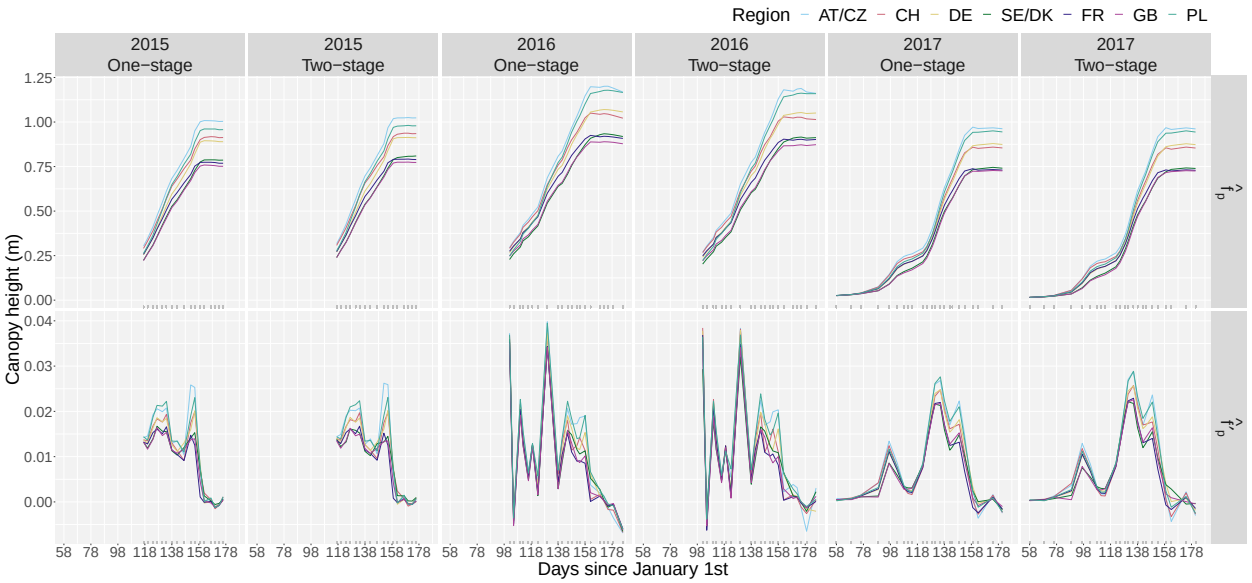


Figure 7.21: For the FIP platform: Region-specific growth curves, $\hat{f}_p$, and region-specific first-order derivatives, $\hat{f}'_p$, by trial (2015, 2016 and 2017) and for the one- and two-stage approaches. AT/CZ: Austria/Czechia; CH: Switzerland; DE: Germany; FR: France; GB: Great Britain; PL: Poland; SE/DK: Sweden/Denmark.

Chapter 8

Software developments

This thesis proposes different P-spline-based models to analyse spatio-temporal HTP data. This chapter discusses and describes the software implementations that allow the practical use of these models. In particular, for the first stage of the two-stage approach described in Chapter 4, we specifically use the R-packages `SpATS` and `statgenHTP` (functions `time points()`, `fitModels()` and `getCorrected()`) to estimate the spatial model (4.2) at different measurement times (see Section 4.1.5). To estimate the longitudinal and hierarchical psHDM (4.12) in the second stage of the two-stage approach (see Section 4.2.5), we created new functions: `fitSplinesModels()`, `predict.psHDM()`, `plot.psHDM()` and `estimateSplineParameters()`, and integrated them to the `statgenHTP()` package. Regarding the one-stage approach presented in Chapter 5, we provide the user with one additional function, `fit3DSplineHDM()`, to estimate the spatio-temporal psHDM (5.2) (see Section 5.3), which is publicly available at https://gitlab.bcamath.org/dperez/htp_one_stage_approach). In what follows, we show the functionalities of the code available and illustrate its usage by reproducing the results for the PhenoArch data presented in Section 7.1. Data are available as "Supplementary Information" for the Perez-Valencia et al. (2022) paper at https://www.nature.com/articles/s41598-022-06935-9#Sec17.

8.1 statgenHTP R-package

The statgenHTP R-package (High Throughput Phenotyping (HTP) Data Analysis; Millet et al., 2022) is part of a series of packages developed by the Biometris research group (Wageningen University & Research) and collaborators to contribute to the knowledge transfer of specialised statistical methods in the context of plant breeding experiments. statgenHTP (see a short overview in https://biometris.github.io/statgenHTP/articles/Overview_HTP.html) provides a set of functions to

1. prepare and describe the data (tutorial 1, https://biometris.github.io/statgenHTP/articles/vignettesSite/Intro_HTP.html),
2. detect outliers at the time point or at the plant levels (tutorials 2 and 4, https://biometris.github.io/statgenHTP/articles/vignettesSite/OutlierSingleObs_HTP.html and https://biometris.github.io/statgenHTP/articles/vignettesSite/OutlierSerieObs_HTP.html),
3. accurately separate the genetic effects from the spatial effects at each time point (tutorial 3, https://biometris.github.io/statgenHTP/articles/vignettesSite/SpatialModel_HTP.html),
4. model the temporal evolution of the genetic signal (tutorial 5, https://biometris.github.io/statgenHTP/articles/vignettesSite/HierarchicalDataModel_HTP.html), and
5. estimate relevant parameters from modelled time courses (tutorial 6, https://biometris.github.io/statgenHTP/articles/vignettesSite/ParameterEstimation_HTP.html).

As part of this thesis, we specifically collaborated with the understanding of some functions for the tutorial 3 (first stage of the two-stage approach), developed functions for the tutorial 5 (second stage of the two-stage approach), and extended the functionalities of functions in the tutorial 6 (extraction of time-independent attributes to characterise genotypes).

8.2 Two-stage approach R-functions

We start this section by introducing the data structure used for this kind of analysis. To illustrate, we use the PhenoArch data, presented in Section 2.1 and later analysed in section 7.1

```
library(statgenHTP)
head(PhenoArchData)
```

```
  timeNumber  timePoint rowId colId plotId    TrtGeno  TrtPanel    LeafArea
1        103 2017-04-13    26    24 c24r26 WD_GenoA01 WD_Panel1 0.003377735
2        103 2017-04-13    56    21 c21r56 WD_GenoA01 WD_Panel1 0.002489031
3        103 2017-04-13     3    28  c28r3 WD_GenoA01 WD_Panel1 0.002515396
```

```
4         103 2017-04-13    24    20 c20r24 WD_GenoA01 WD_Panel1 0.003256119
5         103 2017-04-13     2    16  c16r2 WD_GenoA01 WD_Panel1 0.002871676
6         103 2017-04-13    24    21 c21r24 WD_GenoA02 WD_Panel1 0.002923092
```

`PhenoArchData` is a data frame of dimension 51088×8 (including missing data). The column `timeNumber` corresponds to the DOYs related with the `timePoint` column ($n = 31$). The columns `rowId` and `colId` indicate the row ($R = 60$) and column ($C = 28$) positions of each plant (`plotId`, with $M = 1648$) in the grid. `TrtGeno` refers to the genotypes ($L = 180$, genotype by water regime combination), and `TrtPanel` are the populations ($K = 4$, panel by water regime combination). Finally, the phenotype of interest is the leaf area (`LeafArea`).

Results for the two-stage approach can be obtained with the functions in the `statgenHTP` R-package. In the first stage, we fit the SpATS model (4.2) at each measurement time. To do so, we first need to create an object of the class `TP` (see `help(createtime points)`), i.e., a list of standarised data frames where each one contains the data for a single time point,

```
> PhenoTP <- createTimePoints(dat = PhenoArchData,
+                             experimentName = "PhenoArch",
+                             genotype = "TrtGeno",
+                             timePoint = "timePoint",
+                             plotId = "plotId",
+                             rowNum = "rowId", colNum = "colId")
> summary(PhenoTP)
PhenoTP contains data for experiment PhenoArch.

It contains 31 time points.
First time point: 2017-04-13
Last time point: 2017-05-13

No check genotypes are defined.
```

With the above code, we are indicating where the different experiment information is located in the data frame. Using object `PhenoTP`, we now can fit the SpATS model (with genotypes either as random or fixed effects) at each measurement time point by using the `fitModels()` function. We note that with this function two engines can be used: `SpATS` (Rodríguez-Álvarez et al., 2018) or `ASReml` (Butler et al., 2018). We focus here in the SpATS engine development. The usage is as follows

```
> # Spatial model using SpATS (genotype as random)
> modPheno.ran <- fitModels(TP = PhenoTP,
+                           trait = "LeafArea",
+                           geno.decomp = c("TrtPanel"),
```

```
+                           what = "random")
> summary(modPheno.ran)
Models in modPheno.ran where fitted for experiment PhenoArch.

It contains 31 time points.
The models were fitted using SpATS.

> # Spatial model using SpATS (genotype as fixed)
> modPheno.fix <- fitModels(TP = PhenoTP,
+                           trait = "LeafArea",
+                           what = "fixed")
> summary(modPheno.fix)
Models in modPheno.fix where fitted for experiment PhenoArch.

It contains 31 time points.
The models were fitted using SpATS.
```

The previous code let us fit the SpATS model with genotypes as random (`modPheno.ran` object and argument `what = "random"`) or fixed effects (`modPheno.fix` object and argument `what = "fixed"`). Notice that if genotypes are modelled as random effects, we then can use the `geno.decom` option, which allows considering a different genotypic variances according to the variable specified (in this case the population, `TrtPanel`). Additional fixed effects can also be included into the model through the argument `extraFixedFactors` (e.g., for the FIP data, we used as extra fixed factor the lot effect).

We notice that the `fitModels()` function uses by default the PS-ANOVA formulation for the spatial term $f_S(r,c)$ (see Section 3.2.2). This term is constructed using the function `PSANOVA()` from the `SpATS` package as

```
PSANOVA(colNum, rowNum, nseg = nSeg, nest.div = c(2,2), center = TRUE)
```

where `nSeg = c(number of columns, number of rows)` and `nest.div = c(2,2)` indicates that nested bases, with half the dimension in both directions (rows and columns), are used for the smooth-by-smooth interaction component in $f_S(r,c)$ (see equation (3.26) and Lee et al., 2013). However, for the results in Section 7.1, we slightly modify this function (not shown here but available in https://gitlab.bcamath.org/dperez/htp_one_stage_approach) to set $b_2 = b_3 = 8$ (i.e., `nSeg = c(5,5)` and `nest.div = c(1,1)`) for later fair comparison with the one-stage approach results.

After fitting the SpATS model by time point, we obtain the data frame with the spatially corrected leaf area (see equation (4.4)) by using the `getCorrected()` function. This is the "new" data that is used as input for the second stage of the two-stage approach

```
# Extract longitudinal results from SpATS
# Spatially corrected phenotypic values (genotype as random)
> Pheno.cor.ran <- getCorrected(modPheno.ran)
> str(Pheno.cor.ran)
'data.frame': 51088 obs. of  10 variables:
 $ timeNumber   : int  1 1 1 1 1 1 1 1 1 1 ...
 $ timePoint    : POSIXct, format: "2017-04-13" "2017-04-13" ...
 $ LeafArea_corr: num  NA NA 0.00259 NA 0.00306 ...
 $ LeafArea     : num  NA NA 0.00287 NA 0.00326 ...
 $ wt           : num  5111567 5111567 5111567 5111567 5111567 ...
 $ genotype     : Factor w/ 180 levels "WD_GenoA01","WD_GenoA02",..: 1 1 ...
 $ geno.decomp  : Factor w/ 4 levels "Panel 1 - WD",..: 1 1 1 1 1  ...
 $ rowId        : Factor w/ 60 levels "1","2","3","4",..: 20 6 2 19 24 56 ...
 $ colId        : Factor w/ 28 levels "1","2","3","4",..: 12 13 16 1 ...
 $ plotId       : Factor w/ 1648 levels "c10r1","c10r10",..: 131 233 ...

# Spatially corrected phenotypic values (genotype as fixed)
> Pheno.cor.fix <- getCorrected(modPheno.fix)
> str(Pheno.cor.fix)
'data.frame': 51088 obs. of  9 variables:
 $ timeNumber   : int  1 1 1 1 1 1 1 1 1 1 ...
 $ timePoint    : POSIXct, format: "2017-04-13" "2017-04-13" "2017-04-13" ...
 $ LeafArea_corr: num  NA NA 0.00257 NA 0.00306 ...
 $ LeafArea     : num  NA NA 0.00287 NA 0.00326 ...
 $ wt           : num  5008998 5008998 5008998 5008998 5008998 ...
 $ genotype     : Factor w/ 180 levels "WD_GenoA01","WD_GenoA02",..: 1 1 ...
 $ rowId        : Factor w/ 60 levels "1","2","3","4",..: 20 6 2 19 24 56 ...
 $ colId        : Factor w/ 28 levels "1","2","3","4",..: 12 13 16 1 ...
 $ plotId       : Factor w/ 1648 levels "c10r1","c10r10",..: 131 233 367 ...
```

This new data (`Pheno.cor.ran` and `Pheno.cor.fix`) have two additional columns (regarding the original one, `PhenoArchData`): `LeafArea_corr` (i.e., the spatially corrected leaf area) and `wt` (i.e., the weights used to propagate error from the first to the second stage of the two-stage approach, see equation (4.5)). The `geno.decomp` column is missing in the `Pheno.cor.fix` data because there is no decomposition of the genotypic variance associated with the population effect. Note that the `timeNumber` column returned by the `getCorrected` function is a simple enumeration of the `timePoint` column. Care must be taken when dealing with non-equidistant time points (e.g., the FIP data) to keep the same time distance as in the original `timePoint` column. In this example, we recalculate the `timeNumber` column with time in days of the year (DOYs), that is

```
Pheno.cor.ran$timeNumber <- as.numeric(strftime(Pheno.cor.ran$timePoint, format = "%j"))
Pheno.cor.fix$timeNumber <- as.numeric(strftime(Pheno.cor.fix$timePoint, format = "%j"))
```

For more details on the correction for spatial trends, we refer the reader to the statgenHTP tutorial 3 (https://biometris.github.io/statgenHTP/articles/vignettesSite/SpatialModel_HTP.html). For the second stage of the two-stage approach, we have developed a set of functions to fit (`fitSplineHDM`()), predict (`predict.psHDM()`), plot (`plot.psHDM()`) and extract attributes (`estimateSplineParameters()`) from the psHDM (4.6). Functions to fit and predict are supported with other (hidden) subroutines collected in the source `fitSplineHDMHelperFunctions`. The notation in the R-documentation of these functions is based on the paper by Perez-Valencia et al. (2022). In what follows, we illustrate their usage with the PhenoArch data. For more details on the analysis of the temporal evolution of the genetic signal and further extraction of time-independent attributes with these functions, we refer the reader to the statgenHTP tutorials 5 and 6 (https://https://biometris.github.io/statgenHTP/articles/vignettesSite/HierarchicalDataModel_HTP.html and https://biometris.github.io/statgenHTP/articles/vignettesSite/ParameterEstimation_HTP.html).

8.2.1 `fitSplineHDM` function

We use the `Pheno.cor.ran` data frame, previously obtained in the first stage of the two-stage approach, to fit the psHDM as follows

```
fit.psHDM <- fitSplineHDM(inDat = Pheno.cor.ran,
                          genotypes = NULL,
                          plotIds = NULL,
                          trait = "LeafArea_corr",
                          useTimeNumber = TRUE, timeNumber = "timeNumber",
                          pop = "geno.decomp",
                          genotype = "genotype",
                          plotId = "plotId",
                          weights = "wt",
                          difVar = list(geno = FALSE, plot = FALSE),
                          smoothPop = list(nseg = 8, bdeg = 3, pord = 2),
                          smoothGeno = list(nseg = 8, bdeg = 3, pord = 2),
                          smoothPlot = list(nseg = 8, bdeg = 3, pord = 2),
                          trace = FALSE)
```

In the example above, we fit the psHDM to the spatially corrected leaf area (`trait = "LeafArea_corr"`) as a function of time (we use the numerical time, i.e., the DOYs, as indicated in the column `timeNumber` of the `Pheno.cor.ran` data frame). We assume a hierarchical data structure, with plants (`plotId = "plotId"`) nested in genotypes (`genotype = "genotype"`) and genotypes

nested in populations (`pop = "geno.decomp"`). We use cubic (`bdeg = 3`) B-spline bases of dimension $b_{\text{pop}} = b_{\text{gen}} = b_{\text{plant}} = 11$ and second-order penalties (`pord = 2`) to represent f_p, f_g and f_i in (4.7) as in the data analysis, Section 7.1. We note that the `fitSplineHDM()` function uses as argument the number of segments `nseg` instead of the number of B-spline basis functions, b (`nseg` = $b-$ `bdeg`, that is, for our example, if $b = 11$, then `nseg` = 8). We use the weights obtained after the spatial correction is performed in the first stage (`weights = "wt"`) to propagate the error to the second stage. With the `difVar` argument, the user can also specify if the genetic variation varies across populations (`geno = TRUE`) and the plant variation changes across genotypes (`plot = TRUE`), as proposed in Section 4.2.3. Consequently, the number of variance components will increase with the number of populations and/or genotypes, while the number of coefficients will remain the same. If `trace = TRUE`, a report with changes in deviance and effective dimensions is printed by iteration. It is helpful to detect convergence problems. Finally, the fitting process can also be performed for a subset of genotypes or plots. The user only needs to specify the desired respective vectors in the `genotypes` and/or `plotIds` arguments of the function; in this example, we use the information for all plants and genotypes (i.e., `genotypes = NULL, plotIds = NULL`).

The resulting object, in this case `fit.psHDM`, contains the following information

```
> names(fit.psHDM)
 [1] "y"           "time"        "popLevs"     "genoLevs"    "plotLevs"
 [6] "nPlotPop"    "nGenoPop"    "nPlotGeno"   "MM"          "ed"
[11] "vc"          "phi"         "coeff"       "deviance"    "convergence"
[16] "dim"         "family"      "Vp"          "smooth"      "popLevel"
[21] "genoLevel"   "plotLevel"
```

- Information about the raw data structure (in this case about `Pheno.cor.ran`), e.g., `fit.psHDM$popLevs`, `fit.psHDM$genoLevs` and `fit.psHDM$plotLevs` are factors with the names of the populations, genotypes and plants, respectively.

```
> fit.psHDM$popLevs
[1] Panel 1 - WD Panel 1 - WW Panel 2 - WD Panel 2 - WW
Levels: Panel 1 - WD Panel 1 - WW Panel 2 - WD Panel 2 - WW
```

- Information about the fitting process, e.g., `fit.psHDM$vc` and `fit.psHDM$ed` are, respectively, numeric vectors with the (REML) variance component estimates and associated effective dimensions (or effective degrees of freedom) for each random component of the model (one for each population level and for intercept, slope and non-linear trend at genotype and plant levels).

```
> fit.psHDM$ed
         p1          p2          p3          p4       g.int       g.slp    g.smooth
   8.749409    8.699878    8.598307    8.046305  167.205526  167.827410  834.213007
      i.int       i.slp    i.smooth
1266.142602 1417.752826 3692.254299
```

- Three data frames with the estimated curves at each of the three-level hierarchy (population: `fit.psHDM$popLevel`, genotypes: `fit.psHDM$genoLevel` and plants: `fit.psHDM$plotLevel`).

```
> head(fit.psHDM$popLevel)
  timeNumber  timePoint          pop        fPop   fPopDeriv1   fPopDeriv2
1        103 2017-04-13 Panel 1 - WD 0.002726544 1.179289e-05 1.130232e-03
2        104 2017-04-14 Panel 1 - WD 0.003251929 9.874524e-04 8.210871e-04
3        105 2017-04-15 Panel 1 - WD 0.004598401 1.653967e-03 5.119424e-04
4        106 2017-04-16 Panel 1 - WD 0.006456815 2.011337e-03 2.027977e-04
5        107 2017-04-17 Panel 1 - WD 0.008519722 2.079909e-03 5.642507e-05
6        108 2017-04-18 Panel 1 - WD 0.010684835 2.307306e-03 3.983686e-04
```

  That is, for the example above, the estimated population trajectories ($\hat{f}_p$, `fPop`) and their first ($\hat{f}'_p$, `fPopDeriv1`) and second-order derivatives ($\hat{f}''_p$, `fPopDeriv1`).

For a detailed description of the returned values see `help(fitSplineHDM)`.

8.2.2 `predict.psHDM` function

We use the `fit.psHDM` object to predict the psHDM as follows

```
pred.psHDM <- predict(object = fit.psHDM,
                      newtimes = seq(min(fit.psHDM$time[["timeNumber"]]),
                                     max(fit.psHDM$time[["timeNumber"]]),
                                     length.out = 100),
                      pred = list(pop = TRUE, geno = TRUE, plot = TRUE),
                      se = list(pop = TRUE, geno = TRUE, plot = FALSE),
                      trace = FALSE)
```

In the code above, we use the `fit.psHDM` object to make predictions at the three levels of the hierarchy (`pred = list(pop = TRUE, geno = TRUE, plot = TRUE)`) and to obtain standard errors at the population and genotype levels (`se = list(pop = TRUE, geno = TRUE, plot = FALSE)`). The original

data is measured at 31 time points, but predictions are obtained at 100 time points in the same range as the original time points (argument `newtimes`). We note that standard errors (argument `se`) at the plant level are not calculated (`plot = FALSE`) due to the intensive computing memory and time that could be taken for this calculation. As such, if it is not strictly necessary, we suggest the user set the standard errors at the plot level as `FALSE`.

As a result, three data frames with predictions (and standard errors) at population (`pred.psHDM$popLevel`), genotype (`pred.psHDM$genoLevel`) and plant (`pred.psHDM$plotLevel`) levels are returned. When predictions are calculated on a denser grid of time points, an additional data frame (`pred.psHDM$plotObs`) with the raw data will be returned; otherwise, the data frame at plant level (`pred.psHDM$plotLevel`) will have an additional column (`obsPlot`) with the raw data.

```
> names(pred.psHDM)
[1] "newtimes"  "popLevel"  "genoLevel" "plotLevel" "plotObs"
```

As example, we show a portion of the data frame at the population level (without the `timePoint` column for brevity).

```
> head(pred.psHDM$popLevel[,-c(2)])
  timeNumber          pop       fPop fPopDeriv1 fPopDeriv2      sePop  sePopDeriv1  sePopDeriv2
1   103.0000 Panel 1 - WD 0.00272654 0.00001179 0.00113023 0.00093024 0.0002376660 3.646313e-05
2   103.3030 Panel 1 - WD 0.00278058 0.00034009 0.00103655 0.00092015 0.0002353946 3.316423e-05
3   103.6061 Panel 1 - WD 0.00292979 0.00064001 0.00094287 0.00091540 0.0002334352 3.007650e-05
4   103.9091 Panel 1 - WD 0.00316559 0.00091153 0.00084919 0.00091596 0.0002316798 2.727176e-05
5   104.2121 Panel 1 - WD 0.00347937 0.00115467 0.00075551 0.00092172 0.0002300366 2.484603e-05
6   104.5152 Panel 1 - WD 0.00386252 0.00136942 0.00066183 0.00093249 0.0002284320 2.291997e-05
```

8.2.3 `plot.psHDM` function

This plot function provides five plot types (`plotType`) for objects of the class `psHDM` after fitting (`fitSplineHDM()`) or predicting (`predict.psHDM()`). To illustrate the usage of function `plot.psHDM()`, we use the object `pred.psHDM` obtained above. We can plot at each hierarchy level. We start by showing how to obtain plots at population level

```
## Population-specific trajectories.
plot(pred.psHDM, plotType = "popTra",
     xlab = "DOY", ylab = expression(tilde(y)[i](t)), themeSizeHDM = 20)
```

If `plotType = "popTra"`, estimated population-specific trajectories are depicted ($\hat{f}_p(t)$) separately for each population (see Figure 8.1), and their 95% pointwise confidence intervals. Additionally, the grey lines represent the observed trait that is used in the `fitSplineHDM()` function (i.e., $\tilde{y}_i$, the spatially corrected leaf area).

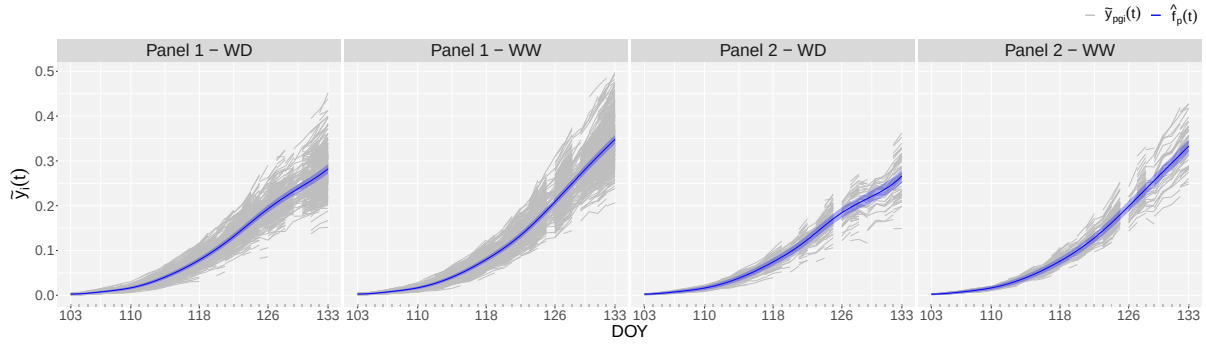


Figure 8.1: For the PhenoArch platform: Estimated population-specific trajectories (blue lines) with 95% pointwise confidence intervals (blue shaded areas). The grey lines represent the spatially corrected leaf area at the plant level (first stage).

At genotype level we can visualise three plots

```
## Population and genotype-specific trajectories.
plot(pred.psHDM, plotType = "popGenoTra",
     xlab = "DOY", ylab = expression(tilde(y)[i](t)),
     themeSizeHDM = 20)

## First-order derivative of the population- and genotype-specific trajectories.
plot(pred.psHDM, plotType = "popGenoDeriv",
     xlab = "DOY", themeSizeHDM = 20)

## Genotype-specific deviations.
plot(pred.psHDM, plotType = "genoDev",
     xlab = "DOY", ylab = expression(tilde(y)[i](t)),
     themeSizeHDM = 20)
```

If `plotType = "popGenoTra"` (see Figure 8.2(a)), estimated population ($\hat{f}_p(t)$) and genotype-specific ($\hat{f}_p(t) + \hat{f}_g(t)$) trajectories are depicted for all genotypes separately for each population. Additionally, 95% pointwise confidence intervals are depicted for the estimated population trajectories. If `plotType = "popGenoDeriv"` (see Figure 8.2(b)), first-order derivative of the estimated population ($\hat{f}'_p(t)$) and genotype-specific ($(\hat{f}_p(t) + \hat{f}_g(t))'$) trajectories are depicted for all genotypes separately for each population, and 95% pointwise confidence intervals are depicted for estimated trajectories at the population level.

Finally, if `plotType = "GenoDev"` (see Figure 8.2(c)), estimated genotype-specific deviations ($\hat{f}_g(t)$) are depicted for all genotypes separately for each population.

We note that Figure 8.2 and Figure 7.5 (except (d)) are equivalent for the two-stage approach case. However, estimated values in Figure 8.2 are calculated on a denser grid of time points and we additionally present here 95% pointwise confidence intervals at population and genotype levels.

We finish the examples of the `plot.psHDM()` function by presenting the usage at the plant level

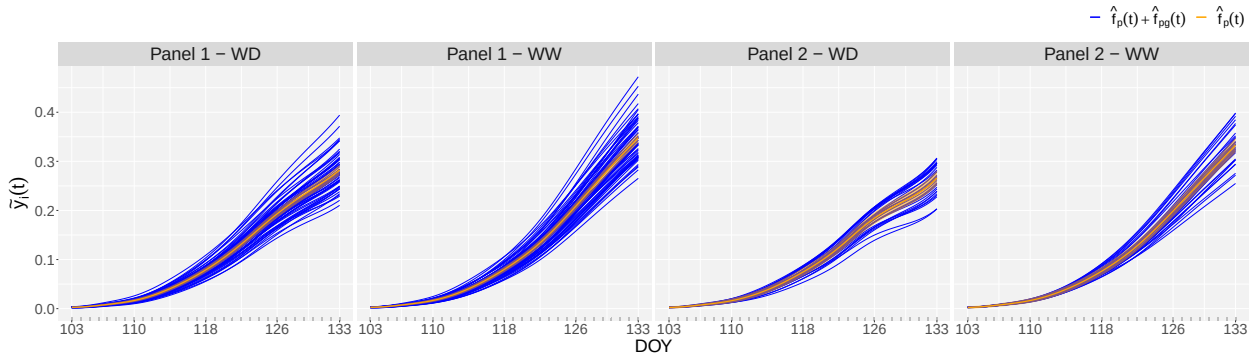
```
## Genotype- and plot-specific trajectories.
## As an example we used the same four genotypes used in Figure 7.5(d)
plot.genos  <- c("WD_GenoA44", "WW_GenoA44", "WD_GenoB20", "WW_GenoB20")
names.genos <- c("Geno 44 - Panel 1 - WD", "Geno 20 - Panel 2 - WD",
                 "Geno 44 - Panel 1 - WW", "Geno 20 - Panel 2 - WW")
plot(pred.psHDM,
     plotType = "genoPlotTra",
     genotypes = plot.genos, genotypeNames = names.genos,
     genotypeOrder = c(1,3,2,4),
     xlab = "DOY", ylab = expression(tilde(y)[i](t)),
     themeSizeHDM = 20)
```

If `plotType = "genoPlotTra"` (see Figure 8.3), estimated genotype ($\hat{f}_p(t) + \hat{f}_g(t)$) and plant-specific ($\hat{f}_p(t) + \hat{f}_g(t) + \hat{f}_i(t)$) trajectories are depicted for all plants separately for a selection of `genotypes`. Also, 95% pointwise confidence intervals are depicted for the estimated genotype-specific trajectories. For this `plotType`, the user has the option to change the names (`genotypeNames`) and/or the order (`genotypeOrder`) of the selected genotypes.

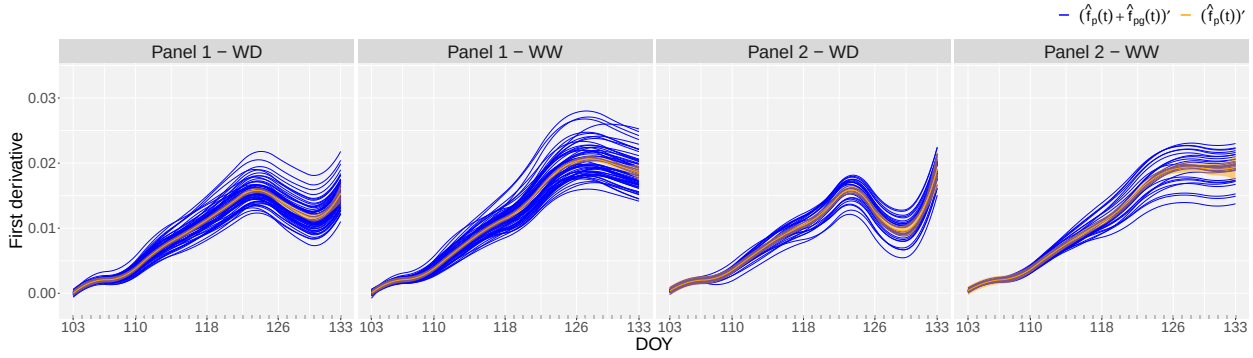
As for figures at the genotype level, Figure 8.3 and Figure 7.5(d) are equivalent for the two-stage approach case. Again, the difference is that estimated values in Figure 8.3 are calculated on a denser grid of time points, and we additionally calculate 95% pointwise confidence intervals at the genotype level. Besides, the grey lines in Figure 7.5(d) represent the raw leaf area, while in Figure 8.3 they represent the spatially corrected leaf area.

8.2.4 `estimateSplineParameters` function

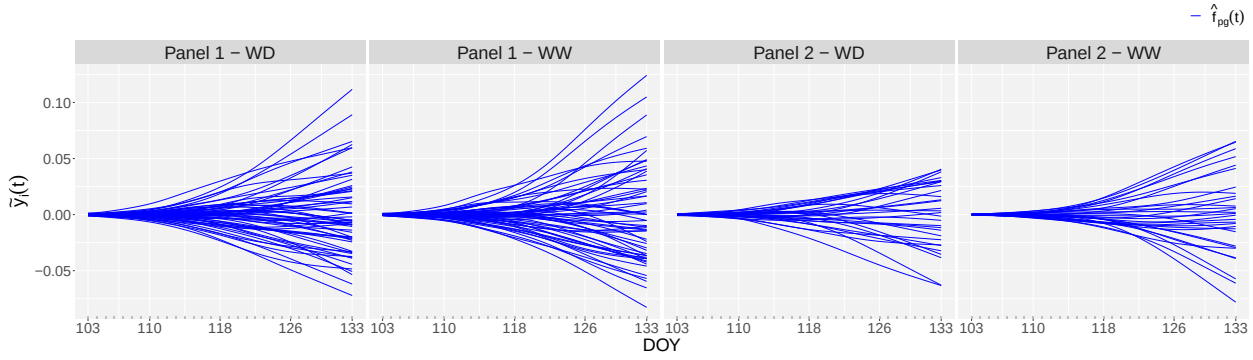
This function extracts parameter estimates from fitted splines on a specified interval. It can be used with class objects obtained with the different methods in the `statgenHTP` package. For this example, we will focus on curves obtained from the P-splines hierarchical data model. That is, we can use objects obtained from `fitSplineHDM()` (fitted curves in the `fit.psHDM` object) or `predict.psHDM()` (predicted curves in the `pred.psHDM` object) functions. We use in this example the predicted curves. Although we have available



(a) Estimated population- and genotype-specific trajectories



(b) Estimated population- and genotype-specific first-order derivatives



(c) Estimated genotype-specific deviations

Figure 8.2: For the PhenoArch platform: (**a**) Estimated population (orange lines) and genotype (blue lines) specific trajectories with 95% pointwise confidence intervals at population level (orange shaded areas), (**b**) Estimated population (orange lines) and genotype (blue lines) specific first-order derivatives with 95% pointwise confidence intervals at population level (orange shaded areas), and (**c**) estimated genotype-specific deviations for all genotypes.

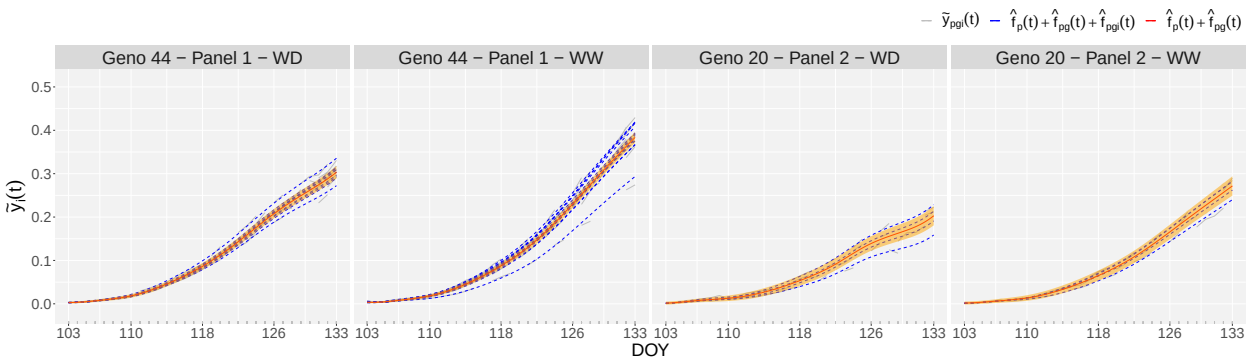


Figure 8.3: For the PhenoArch platform: For the four genotypes used in Figure 7.5(d) (as illustration), estimated genotype (red lines) and plant (dotted blue lines) specific trajectories with 95% pointwise confidence intervals (red shaded areas) at genotype level. The grey lines represent the spatially corrected leaf area at the plant level (first stage).

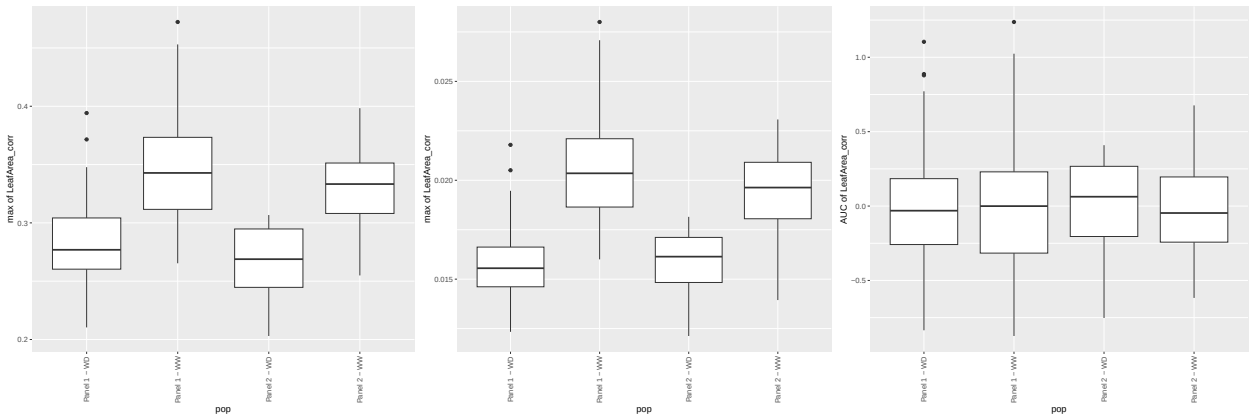
information at population, genotype and plant levels, this function only extracts information at genotype and plant levels. However, we are generally interested in the genotype level. For instance, in Section 7.1.3, we extracted three features: maxTrait, maxSpeed and AUC. We use the function `plot.splineEst()` to plot the boxplots with the estimates in Figure 8.4. The code is as follows

```
## Estimate maximum spatially corrected leaf area.
paramArch1 <- estimateSplineParameters(x = pred.psHDM,
                                       what = "max",
                                       fitLevel = "geno",
                                       estimate = "predictions")
plot(paramArch1, plotType = "box")

## Estimate maximum speed rate.
## We are interested on a local maximum (before timeNumber 130).
paramArch2 <- estimateSplineParameters(x = pred.psHDM,
                                       what = "max",
                                       fitLevel = "geno",
                                       estimate = "derivatives",
                                       timeMax = 130)
plot(paramArch2, plotType = "box")

## Estimate area under the curve (AUC).
paramArch3 <- estimateSplineParameters(x = pred.psHDM,
                                       what = "AUC",
                                       fitLevel = "genoDev",
                                       estimate = "predictions")
```

```
plot(paramArch3, plotType = "box")
```



(a) maxTrait (b) maxSpeed (c) AUC

Figure 8.4: For the PhenoArch platform: Boxplots with the extracted attributes at the genotype level by population (panel by water regime) (**a**) maxTrait, (**b**) maxSpeed, and (**c**) AUC.

With the code above, we calculate: (i) for each genotype trajectory (`fitLevel = "geno"` and `estimate = "predictions"`) in Figure 8.2(a) the estimated maximum (`what = "max"`) spatially corrected leaf area, i.e., the maxTrait, as depicted in Figure 8.4(a) (see also the Figure 7.7 left); (ii) for each first-order derivative of the estimated genotype-specific trajectory (`fitLevel = "geno"` and `estimate = "derivatives"`) in Figure 8.2(b), the estimated maximum (`what = "max"`) speed rate (maxSpeed) of the spatially corrected leaf area before $t = 130$ (`timeMax = 130`), as depicted in Figure 8.4(b) (see also the Figure 7.7 centre); and (iii) for each genotype deviation curve (`fitLevel = "genoDev"` and `estimate = "predictions"`) in Figure 8.2(c) the AUC (`what = "AUC"`), as shown in Figure 8.4(c) (see also the Figure 7.7 right). Objects `paramArch1` and `paramArch2`, additionally contain information about the time point at which the maximum points occur.

8.3 One-stage approach R-functions

Following the same ideas than for the two-stage approach, for the one-stage approach, we have developed one function to fit HTP data with the (3D) spatio-temporal psHDM (5.2). The function, named `fit3DSplineHDM()`, is available on https://gitlab.bcamath.org/dperez/htp_one_stage_approach). We are still working on the implementation of the previous functions to predict, plot and extract attributes. Function to fit is supported with other subroutines collected in the source `fitSplineHDMHelperFunctions()`. We illustrate their use with the PhenoArch data. All the notation used in these functions are based on the paper by Perez-Valencia et al. (2023).

8.3.1 `fit3DSplineHDM` function

We use the original data set `PhenoArchData` to fit the spatio-temporal psHDM as follows

```
fit.3DpsHDM <- fit3DSplineHDM(response = "LeafArea",
                              time = "timeNumber",
                              pop="TrtPanel", geno="TrtGeno", plant="plotId",
                              col = "colId", row = "rowId",
                              data = PhenoArchData,
                              smooth.3D = list(nseg = c(5,5,5), bdeg = 3, pord = 2),
                              smooth.pop = list(nseg = 8, bdeg = 3, pord = 2),
                              smooth.geno = list(nseg = 8, bdeg = 3, pord = 2),
                              smooth.plant = list(nseg = 8, bdeg=3, pord = 2),
                              smooth.rc = list(nseg = 8, bdeg = 3, pord = 2),
                              trace = FALSE)
```

In the code above, we fit the spatio-temporal psHDM (5.2) to the raw leaf area (`trait = "LeafArea"`) as a function of time (we use the numerical time, i.e., the DOYs, as indicated in the column `timeNumber` of the `PhenoArchData` data frame). We assume a hierarchical data structure, with plants (`plant = "plotId"`) nested in genotypes (`geno = "TrtGeno"`) and genotypes nested in populations (`pop = "TrtPanel"`), and the experimental design effects of rows (`row = "rowId"`) and columns (`col = "colId"`). We use cubic (`bdeg=3`), B-spline basis of dimension $b_{\text{pop}} = b_{\text{gen}} = b_{\text{plant}} = b_{\text{row}} = b_{\text{col}} = 11$ and $b_1 = b_2 = b_3 = 8$, and second-order penalties (`pord=2`) to represent f_p, f_g, f_i, f_r, f_c and f_{ST} in (5.3) as in the data analysis, Section 7.1. Similarly to the `fitSplineHDM()` function, `fit3DSplineHDM()` uses as argument the number of segments `nseg` instead of the number of B-spline basis b (for our example, if $b = 8$, then `nseg` = 5). If `trace = TRUE`, a report with changes in deviance and effective dimension is printed by iteration.

The resulting object, `fit.3DpsHDM`, contains different information about the data structure, the fitting process, and eight lists with data frames with: estimated curves (i.e., estimated trajectories and deviations, as well as their first and second-order derivatives) at each of the three-level hierarchy (population: `fit.3DpsHDM$eta_pop`; genotypes: `fit.3DpsHDM$eta_geno` and `fit.3DpsHDM$eta_geno_dev`; and plants: `fit.3DpsHDM$eta_plant` and `fit.3DpsHDM$eta_plant_dev`), estimated curves for the row and column effects (rows: `fit.3DpsHDM$eta_row`, and cols: `fit.3DpsHDM$eta_col`), and the fitted values (`fit.3DpsHDM$fitted_values`), $\hat{y}_i$.

```
> names(fit.3DpsHDM)
 [1] "y"               "time"            "l.plant"         "l.geno"          "l.pop"
 [6] "n.plants_p_pop"  "n.geno_p_pop"    "n.plants_p_geno" "n.row"           "n.col"
[11] "MM"              "ed"              "tot_ed"          "vc"              "phi"
[16] "coeff"           "eta_pop"         "eta_geno_dev"    "eta_geno"        "eta_plant"
```

```
[21] "eta_plant_dev"   "fitted_values"   "eta_row"         "eta_col"         "spatial"
[26] "deviance"        "convergence"     "dim"             "dim.com"         "family"
[31] "smooth"
```

To convert the lists to data frames (for an easy manipulation for, e.g., plotting), we use the auxiliar function `list.to.df()`

```
> fit.3DpsHDM.df <- list.to.df(object = fit.3DpsHDM)
> names(fit.3DpsHDM.df)
[1] "pop.tra"       "geno.tra"      "plant.tra"     "fitted.values" "plant.obs"
[6] "rows"          "cols"
```

As a result, the object `fit.3DpsHDM.df` contains the above mentioned lists with estimated curves in seven data frames: population curves (`fit.3DpsHDM.df$pop.tra`), genotype curves (`fit.3DpsHDM.df$geno.tra`), plant curves (`fit.3DpsHDM.df$plant.tra` and `fit.3DpsHDM.df$plant.obs` for the raw data), row curves (`fit.3DpsHDM.df$rows`), column curves (`fit.3DpsHDM.df$cols`) and fitted values (`fit.3DpsHDM.df$fitted.values`). For instance, the data frame at the population level would be

```
> head(fit.3DpsHDM.df$pop.tra)
      eta_pop eta_pop_deriv1 eta_pop_deriv2          pop timepoint
1 0.003313289   0.0001389505   1.223948e-03 Panel 1 - WD       103
2 0.004006409   0.0011894837   8.771185e-04 Panel 1 - WD       104
3 0.005576647   0.0018931875   5.302892e-04 Panel 1 - WD       105
4 0.007677174   0.0022500621   1.834599e-04 Panel 1 - WD       106
5 0.009962941   0.0022814615   7.464538e-06 Panel 1 - WD       107
6 0.012304219   0.0024571794   3.439712e-04 Panel 1 - WD       108
```

where `eta_pop` ($\hat{f}_p$) are the estimated population trajectories, and `eta_pop_deriv1` ($\hat{f}'_p$) and `eta_pop_deriv2` ($\hat{f}''_p$) are their first and second-order derivatives. All the resulting curves (at the three hierarchy levels) can be used to compare the results with those obtained previously for the two-stage approach, and to extract attributes for genotype characterisation. Although we are still working on the implementation of functions to plot and extract attributes, here, we reproduce Figure 7.5 for the results with the one-stage approach

```
# Rename datasets
data.raw.plant <- fit.3DpsHDM.df$plant.obs
data.plant     <- fit.3DpsHDM.df$plant.tra
data.geno      <- fit.3DpsHDM.df$geno.tra
```

```
data.pop       <- fit.3DpsHDM.df$pop.tra
# Other plot parameters
ntime <- length(unique(data.raw.plant$timepoint))
min.t <- min(data.raw.plant$timepoint)
max.t <- max(data.raw.plant$timepoint)
p <- ggplot() +
    geom_rug(data = data.raw.plant,
             aes(x = timepoint, y = NULL),
             color = "gray", length = unit(0.01, "npc")) +
    scale_x_continuous(breaks = round(seq(min.t, max.t, length.out = 5), 0)) +
    facet_grid( ~ pop)
# Estimated population- and genotype-specific trajectories
my.cols <- c("1" = "blue", "2" = "orange")
p + geom_line(data = data.geno,
              aes(timepoint, eta_geno, group = geno, color = "1")) +
    geom_line(data = data.pop,
              aes(timepoint, eta_pop, group = pop, color = "2"),
              size = 1) +
    scale_color_manual(values = my.cols,
                       labels = c(expression(hat(bolditalic(f))[p] +
                       hat(bolditalic(f))[g]),
                       expression(hat(bolditalic(f))[p]))) +
    labs(x = "DOY",
         y = expression(paste("Leaf area (",m^2, "plan",t^{-1},")")),
         color = "")
```

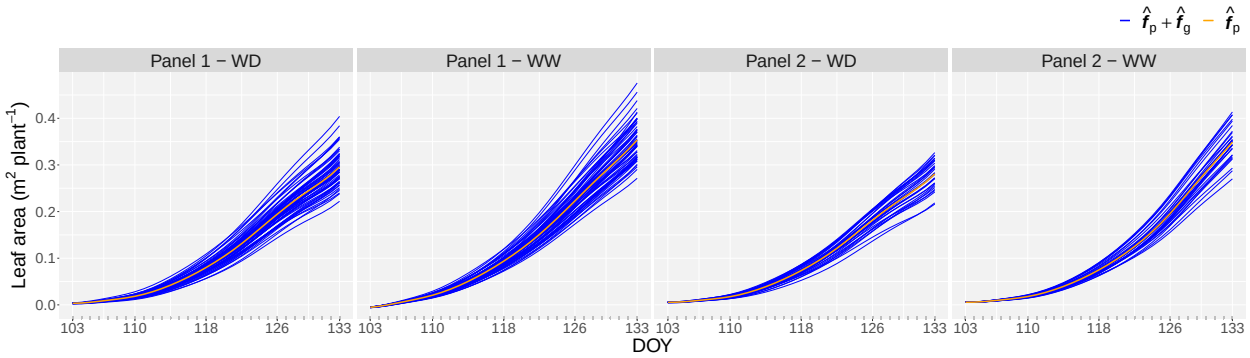


Figure 8.5: For the PhenoArch platform: Estimated population (orange lines) and genotype (blue lines) specific trajectories for all genotypes (see Figure 7.5(a)).

```
# Estimated population and genotype-specific first-order derivatives
p + geom_line(data = data.geno,
              aes(timepoint, eta_geno_deriv1, group = geno, colour = "1")) +
    geom_line(data = data.pop,
              aes(timepoint, eta_pop_deriv1, group = pop, color = "2"),
              size = 1) +
    scale_color_manual(values = my.cols,
                       labels = c(expression(hat(bolditalic(f))*minute[g]),
                                  expression(hat(bolditalic(f))*minute[p]))) +
    labs(x = "DOY",
         y = "First-order derivative",
         color = "")
```

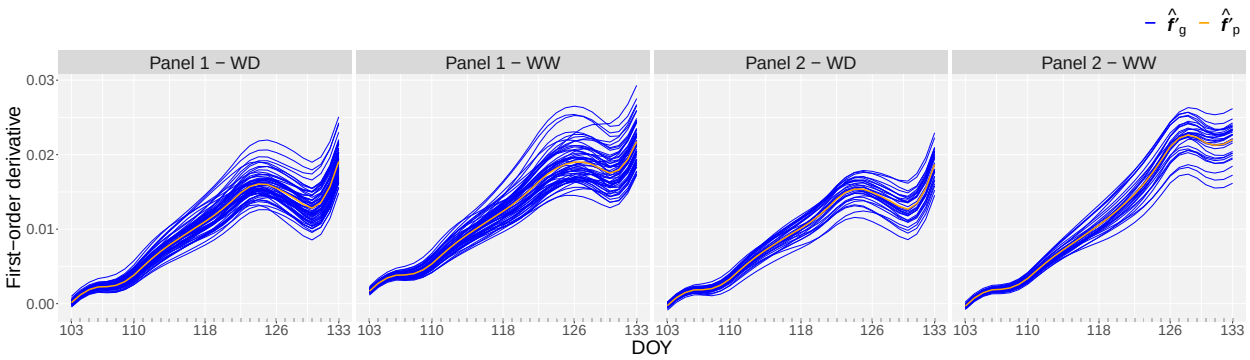


Figure 8.6: For the PhenoArch platform: Estimated population (orange lines) and genotype (blue lines) specific first-order derivatives for all genotypes (see Figure 7.5(b)).

```
# Genotype-specific deviations
ggplot(data = data.geno) +
    geom_line(aes(timepoint, eta_geno_dev, group = geno, colour = "1")) +
    scale_color_manual(values = my.cols,
                       labels = c(expression(hat(bolditalic(f))[g]))) +
    labs(x = "DOY",
         y = expression(paste("Leaf area (",m^2, " plan",t^{-1},")")),
         color = "")
```

```
# Estimated genotype and plant-specific trajectories
  # We choose four genotypes as illustration
  geno.sub       <- c("WD_GenoA44","WD_GenoB20","WW_GenoA44","WW_GenoB20")
  geno.sub.names <- c("Geno 44 - Panel 1 - WD","Geno 20 - Panel 2 - WD",
                      "Geno 44 - Panel 1 - WW","Geno 20 - Panel 2 - WW")
```

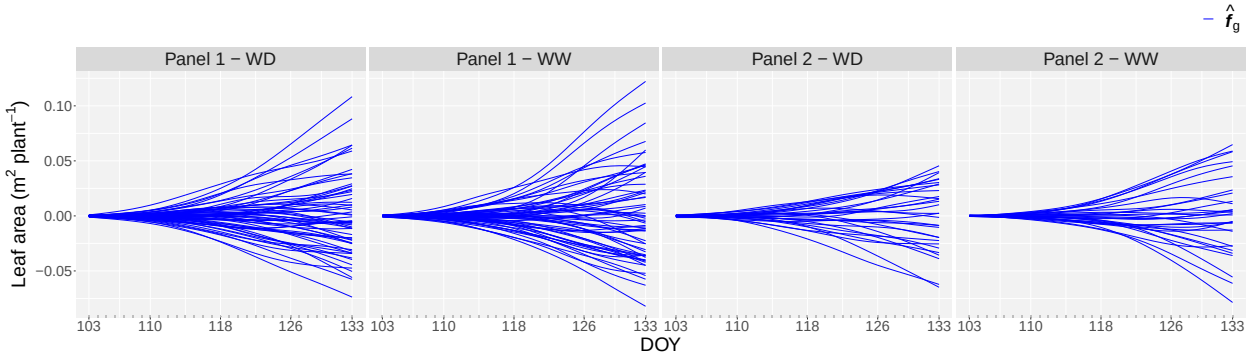


Figure 8.7: For the PhenoArch platform: Estimated genotype-specific deviations for all genotypes (see Figure 7.5(c)).

```
geno.sub.order <- c(1,3,2,4)
# Subset the datasets
data.plant.sub <- droplevels(data.plant[data.plant$geno %in% geno.sub,])
data.raw.sub   <- droplevels(data.raw.plant[data.raw.plant$geno %in% geno.sub,])
data.geno.sub  <- droplevels([$geno %in% geno.sub,])
# Rename the genotypes
data.plant.sub$geno <- factor(data.plant.sub$geno[drop=TRUE], labels = geno.sub.names)
data.raw.sub$geno   <- factor(data.raw.sub$geno[drop=TRUE], labels = geno.sub.names)
data.geno.sub$geno  <- factor(data.geno.sub$geno[drop=TRUE], labels = geno.sub.names)
# Order the genotypes
data.plant.sub$geno <- factor(data.plant.sub$geno,
                              levels = levels(data.plant.sub$geno)[geno.sub.order])
data.raw.sub$geno   <- factor(data.raw.sub$geno,
                              levels = levels(data.raw.sub$geno)[geno.sub.order])
data.geno.sub$geno  <- factor(data.geno.sub$geno,
                              levels = levels(data.geno.sub$geno)[geno.sub.order])
# Plot the desired genotypes
my.cols <- c("1" = "blue", "2" = "gray", "3" = "black")
ggplot() +
  geom_line(data = data.raw.sub,
            aes(timepoint, obs_plant, group = plant, color = "1")) +
  geom_line(data = data.plant.sub,
            aes(timepoint, eta_plant, group = plant,
                linetype = "One-stage", color = "2"),
            show.legend = FALSE) +
  geom_line(data = data.geno.sub,
            aes(timepoint, eta_geno, group = geno,
                linetype = "One-stage", color = "3"),
```

```
            size = 0.8, show.legend = FALSE) +
  geom_rug(data = data.raw.sub,
           aes(x = timepoint, y = NULL),
           color = "gray", length = unit(0.01, "npc")) +
  scale_x_continuous(breaks = round(seq(min.t, max.t, length.out = 5), 0)) +
  scale_color_manual(values = my.cols,
                     labels = c("Raw data",
                                expression(hat(bolditalic(f))[p] +
                                hat(bolditalic(f))[g] +
                                hat(bolditalic(f))[i]),
                                expression(hat(bolditalic(f))[p] +
                                hat(bolditalic(f))[g]))) +
  labs(x = "DOY",
       y = expression(paste("Leaf area (",m^2, " plan",t^{-1},")")),
       color = "") +
  facet_grid( ~ geno)
```

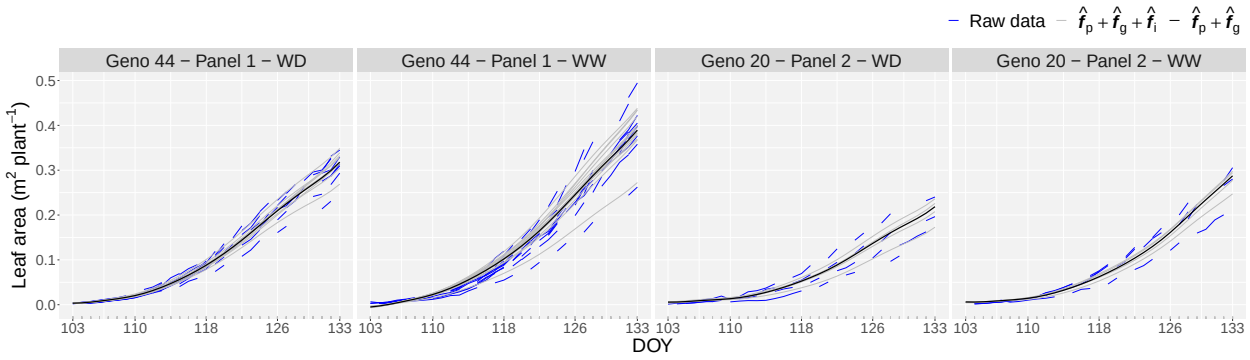


Figure 8.8: For the PhenoArch platform: For the four genotypes used in Figure 7.5(d) (as illustration), estimated genotype (black lines) and plant (grey lines) specific trajectories. The blue lines represent the raw leaf area (see Figure 7.5(d)).

Similarly, to extract time-independent attributes to characterise genotypes (as we did in Section 7.1.3), we use the following code (we specifically obtain three features: maxTrait from Figure 8.5, maxSpeed from Figure 8.6 and AUC from Figure 8.7)

```
# Function to extract features
features <- function(x) {
                      maxTrait   = max(x$eta_geno)
                      xx         = x[x$timepoint < 131, ]
                      maxSpeed   = max(xx$eta_geno_deriv1)
                      Area       = MESS::auc(x$timepoint, x$eta_geno_dev,
```

```
                                        type = "spline", absolutearea = FALSE)
                    pop        = unique(x$pop)
                    geno       = unique(x$geno)
                    res <- data.table::data.table(maxTrait = maxTrait,
                                                  maxSpeed = maxSpeed,
                                                  AUC = Area,
                                                  pop = pop, geno = geno,
                                                  key = c("geno","pop"))}
# It is performed for each genotype
data.geno.s <- split(data.geno, data.geno$geno)
geno.feat   <- lapply(data.geno.s, features)
geno.feat   <- data.table::data.table(do.call("rbind", geno.feat), key = "geno")
```

Once the features are calculated, we plot bivariate scatterplots of the extracted genotype-specific attributes in Figure 8.9. These results for the one-stage approach are the same that we compare with the two-stage approach in Figure 7.7.

```
# ggpairs for extracted features
library(GGally)
ggpairs(data = geno.feat[,-c("geno")],
        aes(color = pop),
        upper = list(continuous = wrap("cor", size = 10)),
        diag = list(continuous = wrap("densityDiag", alpha = 0.6, color = NA)))
```

Figure 8.9: For the PhenoArch platform: Scatterplots matrix with the extracted attributes at the genotype level. The lower off diagonal depicts bivariate scatterplots, the diagonal shows the conditional densities of each attribute per population (panel by water regime combination), the upper off diagonal indicates the bivariate Pearson correlation (marginal and by region; "***" p-value < 0.001, "**" p-value < 0.01, "*" p-value < 0.05, "." p-value < 0.10 and "" otherwise), the last column displays the boxplots of each attribute per population, the last row depicts the conditional histograms of each attribute per population, and the bottom right barplot shows the number of genotypes per population.

Chapter 9

Conclusions

This dissertation "Spatio-temporal modelling of high-throughput phenotyping data" collects the research work done during these last years. Here discuss the contributions, limitations, and challenges that guided the research and outline directions for future work.

HTP data impose a significant challenge due to the complexity/dimensionality and size of the problem. As mentioned in the Introduction, one of the key challenges of this research area includes an in-depth exploration of the methodological path to be followed and its implications. In an effort to pose a comprehensive overview for discussion, in this thesis, we proposed two methodological paths, which compromise the loss of information and the computational complexity. To enhance computational complexity, we first explored decomposing the required spatio-temporal analysis of HTP data into two stages (**Chapter** 4), following the work by Kar et al. (2020), Roth et al. (2021), and van Eeuwijk et al. (2019). In the first stage, we used a two-dimensional spatial model; in the second stage, we used a hierarchical longitudinal model. Subsequently, and to leverage all the information shared by the correlation data structure, we proposed to fit the spatio-temporal effects using a three-dimensional model (**Chapter** 5), in a similar vein to the work by Verbyla et al. (2021). Additionally, to allow for statistical flexibility, we promoted P-splines-based models.

All the details of our two-stage approach are described in **Chapter** 4. In the first stage, we correct the "raw" HTP data for (nuisance) spatial variation and obtain spatially corrected time-series at the resolution of plants or plots with reduced between replicates/plots variability. The second stage consists of a temporal analysis with a hierarchical curve data model to jointly estimate curves at each hierarchy level (population, genotype, and plant or plot) and their first-order derivatives. Apart from this work, van Eeuwijk et al. (2019) also proposed a two-stage approach for analysing HTP data, where they first correct for spatial variation and then focus on estimating and further processing the temporal dynamic of the genetic effects. In that paper, spatially adjusted genotypic means are carried to the temporal analysis. In contrast, our proposal allows keeping the data resolution for the second stage at the experimental unit. Also, in the second stage, we

jointly model the whole sample of spatially corrected curves, while in van Eeuwijk et al. (2019) analyses are done separately per genotype. Our hierarchical approach thus allows borrowing strength across plant/plot curves to more efficiently estimate genotype and population trajectories. This is particularly important in the presence of incomplete data. When choosing a two-stage approach, one might also choose, as done by Roth et al. (2021), to first model the longitudinal variation at a plant or plot level and subsequently apply a spatial correction to extracted features. In that paper, a P-spline model is first fitted separately for each plot time-series, from which the timing of key stages (among other features) are extracted. These intermediate traits are then processed to obtain spatially adjusted genotypic means for further analyses. We feel that both options - with the spatial or the temporal analysis first - represent valid alternatives. The choice for one or another methodological path will depend on the relative magnitudes of the various spatial, temporal and spatio-temporal genetic and non-genetic processes and may be difficult to assess beforehand. A study of the proposal that best suits particular situations represents an interesting area of study. Another important consideration when working with stage-wise approaches is propagating uncertainty from stage to stage. Here, we accomplished this by weighting the second stage with the inverse of the estimated variance associated with the spatially corrected trait. In the HTP context, the weighting has been shown to improve results (Buntaran et al., 2020; Roth et al., 2021).

In **Chapter** 5, we proposed a one-stage spatio-temporal P-spline-based hierarchical approach to model genetic and non-genetic variation in HTP data. From a modelling perspective, the simultaneous modelling of spatial and temporal genetic and non-genetic variation in a one-stage model serves to share information on common spatial variability across measurement times, and therefore overcomes the loss of information given by stage-wise approaches. Yet, one-stage approaches have the limitation of being very computationally demanding, especially when the number of observations and/or the parameters to be estimated is very large. To address this issue, we used the (two-dimensional) SpATS model as the base model and extended it to the (three-dimensional) spatio-temporal case, considering a three-level hierarchical data structure (populations, genotypes within populations, and plants within genotypes). To make our proposals computationally affordable, we combined different specialised methods that take advantage of the sparse model matrices structure (Boer & van Rossum, 2022), the array data structure (Currie et al., 2006), and the non-standard form of the variance-covariance matrix (Rodríguez-Álvarez et al., 2019; Rodríguez-Álvarez et al., 2015). This allowed for efficient computation, even with large datasets.

From both approaches, we obtained estimated curves at the three hierarchy levels and their derivatives, and showed how to calculate from these curves new phenotypic traits, attributes that we called intermediate traits (see Sections 4.2.7 and 5.5). We note that the decision about what summary statistics (intermediate traits such as maximum, minimum, average values and area under de curves) to derive from the models will ultimately depend on the species, the biological phenotypic trait analysed, the applied treatments and/or the range of phenological stages at which the measurements were taken. Although we did not cover it in this thesis explicitly, these new phenotypic traits and the estimated curves can be used for selection purposes

in plant breeding, i.e., to differentiate between genotypes. For instance, estimates for intermediate-level traits can be used as genotypic covariates in models for target traits of commercial interest (e.g., yield and quality parameters), as described by van Eeuwijk et al. (2019). Target traits can be understood as functions of biological (e.g., leaf area for the PhenoArch data and canopy height for the FIP data) and intermediate phenotypic traits for either or both biological and statistical reasons. For instance, yield can be interpreted as a target trait that can be modelled as a function of yield components, where the yield components may represent biological and intermediate level phenotypic traits. For instance, in Roth et al. (2021) intermediate traits obtained from modelling HTP data are included into genotype-by-environment interaction analyses, and Moreira et al. (2020) discuss using information obtained from HTP time-series traits for genomic selection and detecting QTL and causal variants.

An important matter in this thesis was to select and combine appropriate (spatial and longitudinal) statistical methods to be used in our modelling approaches. For the two-stage approach, we decided to use the SpATS model, and hierarchical P-splines were used for the second stage. Conversely, for the one-stage approach, we used a spatio-temporal and hierarchical P-spline-based model. That is, we decided to be consistent throughout our proposals and follow the same modelling philosophy. However, our two-stage approach can be flexible regarding the choice of methods used. For instance, the separable autoregressive model (Gilmour et al., 1997) represents a clear alternative for the spatial component, while for the longitudinal part, hierarchical functional principal component analysis can be used (Xu et al., 2021). We believe that our P-spline-based two-stage approach is attractive both computationally and for interpretation, and the HTP data we analysed in this thesis and other projects show that it works well (see Millet et al., 2022, and https://eppn2020.plant-phenotyping.eu/ for more examples). As far as one-stage approach is concerned, the choice and combination of methods is not straightforward. For instance, Verbyla et al. (2021) combined two models (in only one stage), with one capturing genetic effects through a factor analytic model and the other accounting for spatio-temporal non-genetic residual effects using smoothing splines. They showed that the two models impact each other in the sense that the entire approach might properly represent the correlations present in the data, and then the fitted process must be a simultaneous process rather than independent. Besides, model selection for the involved methods, such as factor analytic models, can introduce complexity and increased computing time to the fitting process. Instead, the authors proposed exploring random regression spline models for genetic effects. In our proposal, we decided to model phenotypic variation as an additive decomposition of the (longitudinal and hierarchical) genetic and (spatio-temporal) non-genetic effects with independent residuals. Although we avoid complications associated with model selection issues, identifiability problems must be carefully addressed.

In **Chapter** 6, we conducted a simulation study to evaluate the performance of our two methods and compare their results. To replicate data commonly encountered in HTP experiments, we proposed a simulated spatio-temporal and hierarchical data structure that enabled us to evaluate results at higher hierarchy levels, such as population and genotype levels. Analyses of the results at each hierarchical level gave us a

broader view of the performance of our methods. Our main findings indicated that, for most simulated situations, the two approaches performed similarly, except for non-nested B-spline bases. Notably, the one-stage approach outperformed the two-stage approach when estimating first-order derivative curves and smaller bases resulted in better performance for both approaches. While our simulation study provided valuable insights, we acknowledge the importance of integrating biological understanding into the problem. In a previous attempt, we simulated data (plant trajectories) by combining statistical-genetic and crop-growth models as described in Bustos-Korts et al. (2019). However, keeping the three-level hierarchical structure was problematic. For this first attempt, additional curves with the average field conditions for each genotype were considered as the "typical genotypic curves", but lacked benchmarks for population curves still persisted. To address the hierarchical data structure issue, we resorted to simulating data from the population to the plant level. Moving forward, it is desirable to develop alternative simulation strategies that consider both statistical and biological perspectives.

Analyses with real HTP data in **Chapter** 7 were also used to evaluate and compare the performance of the one- and two-stage approaches. For the first stage of the two-stage approach, spatially corrected traits with data from the two HTP platforms showed essentially identical results when modelling genotypes as fixed or random effects. Although genotypic fixed effects are recommended for stage-wise approaches (to avoid double-shrinkage), we use random genotypic effects at the first stage for several reasons: (i) genotypes are considered as a random sample, (ii) identifiability problems are avoided (Brumback & Rice, 1998), (iii) random effects improve precision compared to fixed effects (Piepho et al., 2008), (iv) shrinkage of the genotypic BLUPs is counteracted by the inclusion of the residual component into the correction, and (v) heritability can be computed for each measurement time. We propagated the error (through weights), when going from the first to the second stage. Regarding the estimated spatial trends, they varied more smoothly in time when using the one-stage approach. This is a consequence of the fact that for the first stage of the two-stage approach, analysis is performed separately for each time point, and as such, the information on spatial heterogeneity is not shared across time. We obtained estimated curves at the three hierarchy levels as the final results of our approaches. The results at the genotype level follow the simulation results, i.e., minor differences between the two approaches were detected, except for the first-order derivatives for the genotype-specific growth curves. Nevertheless, highly correlated results were obtained between the two approaches when we extracted important time-independent features from these curves, which can be used in posterior analysis (e.g., genotype selection). In our approach, genotypes are modelled as random effects to avoid identifiability problems. However, we acknowledge the potential risk of double-shrinkage in these subsequent analyses with the extracted time-independent characteristics. Using de-regression methods (see, e.g., de Oliveira et al., 2018; Garrick et al., 2009) to obtain unshrunken genotypic BLUPs in our setting would be worth exploring. The FIP experiment (with the three trials) also allowed for an interesting analysis to characterise regional adaptation and to assess genotype consistency across trials with our approaches. Most interestingly, we find a clustering by regions of origin that varies over time, strengthens the potential

of the analysis that can be performed with data collected with these HTP platforms.

In **Chapter** 8 we outlined the functionalities of the code implemented during the course of this project. The implementation of the new advances in the R statistical environment is expected to facilitate the use of the proposed approaches by practitioners and researchers from diverse fields. Our objectives are threefold: to transfer knowledge and technology, to aid in the development of these research areas, and to ensure the reproducibility of scientific results. To make our proposals accessible to practitioners, we made the implemented R-functions publicly available in the `statgenHTP` R-package (https://CRAN.R-project.org/package=statgenHTP, Millet et al., 2022) for data analysis with the two-stage approach. Additionally, we made the code and data required to reproduce the analyses and results of the two-stage approach paper (Perez-Valencia et al., 2022) available at https://gitlab.bcamath.org/dperez/htp_two_stage_approach, and for the one-stage approach paper (Perez-Valencia et al., 2023), at https://gitlab.bcamath.org/dperez/htp_one_stage_approach. We plan to include the one-stage approach functions in the aforementioned package soon to simplify accessibility.

Before proceeding, fairness requires us to mention some limitations of our work. We proposed P-spline-based approaches, which means that as the number of plants and/or B-spline basis dimensions increase, so does the number of parameters to be estimated, and, consequently, the computational time. Typically, computational times are within an acceptable range (in contrast to, e.g., Verbyla et al., 2021). For the experiments analysed in this thesis (PhenoArch and FIP with three trials), estimation for the two-stage approach took around 1 minute and 20 seconds, and for the one-stage approach, computation times were around 25 minutes and one and a half hours without any convergence issues. Nevertheless, the approaches may not scale well to experiments where the number of plants (and associated basis dimension) is very large (due to the size of the system of equations to be solved). Regarding the number of B-spline basis functions, our recommendation (for both approaches) is to use the same value for the three hierarchy levels and the row and column random effects (for the one-stage approach), even if this increases computation. Regarding the number of B-spline basis functions used for the three-dimensional surface (in the row, column and time directions for the one-stage approach), we suggest keeping them relatively small to enable the solution to run on standard computers. The final numbers does not seem to significantly impact results (estimated curves), provided they are large enough to capture the underlying patterns. However, the estimated first-order derivatives have shown to be more sensitive to the number of basis functions.

We finish this thesis by highlighting some opportunities for future work arising from our research. Although we could explore a more general one-stage approach, our current formulation of the spatio-temporal psHDM (5.2) and its implementation (code) is very specific. For instance, we only consider the spatio-temporal smooth function and random effects for rows and columns as non-genetic effects, but the proposal can be extended by taking into account other experimental factors as we did in the first stage of the two-stage approach. Furthermore, while we have focused in this thesis on data with a nested structure, the proposed

modelling framework can be extended to accommodate more complex structures, such as data with crossed levels of grouping (Brumback & Rice, 1998) (e.g., when modelling genotype-by-treatment or genotype-by-environment interactions are of interest). For instance, for the PhenoArch data a crossed-effect structure, to explicitly modelling the genotype-by-treatment interaction, would be more appropriate. Analyses with both simulated and real data have highlighted that the most controversial results are for the first-order derivatives curves. Although derivatives estimation is out of the scope of this thesis, they are important to extract relevant information on genotype performance. Improvements in this area are gaining attention (see, e.g., Hernández et al., accepted, 2023; Simpkin et al., 2018) and show that it is worth exploring in this direction. From a plant breeding perspective, the development of HTP experiments has opened up opportunities to search for new definitions and extensions of the notion of heritability as a function of time. Such definitions will allow for heritability dynamics that are not possible with definitions for traditional experiments (see, e.g., Rodríguez-Álvarez et al., 2018). For instance, Xu et al. (2021) propose a functional measure obtained through HFPCA. Regarding the correlation structure at the genotype level, we assumed the identity matrix times the P-spline matrix for the genotypes as the variance-covariance matrix, but other more interesting variance-covariance structures, such as kinship relationships (instead of the identity matrix) between genotypes, can also be explored (Schmidt et al., 2019; van Eeuwijk et al., 2019). However, we warn of computational complications, even when mathematically it is straightforward. It is worth noting that our current analyses focus on individual trials, and then we believe that the kinship matrix would be better suited for subsequent analyses (e.g., Moreira et al., 2020, for the integration of genetic and phenomic information). Moreover, the results of our approaches (more specifically, the estimated genotype deviations) can be exploited to calculate such kinship matrices. Last but not least, extensions of our approaches would include considering correlations between intercepts and slopes at genotype and plant levels and explicitly imposing constraints on the non-linear/smooth random effect coefficients (Brumback, 2010; Currie, 2014) at genotype (for each population) and plant levels (for each genotype).

All in all, we believe this thesis represents a promising starting point for the analysis of spatio-temporal and hierarchical HTP data. The two proposed approaches represent a good compromise between flexibility, accuracy, adequacy, computational efficiency and interpretability. Our results demonstrate the feasibility of our proposals on standard computers, providing valuable descriptions of the genetic (and non-genetic) variation in the temporal dimension and useful summary statistics for breeding purposes. We believe they represent powerful tools for routine application in phenotyping experiments with dense time series. In our experience, obtaining results from stage-wise approaches is computationally simpler. However, one-stage approaches will always represent a fully efficient choice (see, e.g., Damesa et al., 2017; Schulz-Streeck et al., 2013) since they simultaneously incorporate all sources of variation in a single model. This also means that one-stage approaches can fully account for the variance-covariance structure of the observed data, avoiding substantial loss of valuable information (e.g. spatial heterogeneity across time). Besides, in our work with the analysis of different HTP data, we have observed that the one-stage approach performs better in the

presence of missing data (since it borrows strength across plant curves; more exploration in this direction is required using simulated data). Although the size and complexity/dimensionality of the data and models in this framework are a big challenge for single-stage approaches, our proposal is not only competitive but also novel, providing the groundwork for more sophisticated models. We believe a good practice would be to use a two-stage approach as a starting point to establish the basis for a model in one stage.

References

Anderegg, J., Yu, K., Aasen, H., Walter, A., Liebisch, F., & Hund, A. (2020). Spectral vegetation indices to track senescence dynamics in diverse wheat germplasm. *Frontiers in Plant Science*, *10*, 1749. https://doi.org/10.3389/fpls.2019.01749

Andrade, M. H. M. L., Fernandes Filho, C. C., Fernandes, M. O., Bastos, A. J. R., Guedes, M. L., Marcal, T. d. S., Gonçalves, F. M. A., Pinto, C. A. B. P., & Zotarelli, L. (2020). Accounting for spatial trends to increase the selection efficiency in potato breeding. *Crop Science*. https://doi.org/10.1002/csc2.20226

Araus, J. L., & Cairns, J. E. (2014). Field high-throughput phenotyping: The new crop breeding frontier. *Trends in Plant Science*, *19*(1), 52–61. https://doi.org/10.1016/j.tplants.2013.09.008

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bates, D., Maechler, M., & Jagan, M. (2022). *Matrix: Sparse and dense matrix classes and methods* [R package version 1.5-1]. https://CRAN.R-project.org/package=Matrix

Besag, J., & Higdon, D. (1999). Bayesian analysis of agricultural field experiments. *Journal of the Royal Statistical Society: Series B*, *61*(4), 691–746. https://doi.org/10.1111/1467-9868.00201

Boer, M. P. (accepted, 2023). Tensor product P-splines using a sparse mixed model formulation. *Statistical Modelling*.

Boer, M. P., & van Rossum, B.-J. (2022). *LMMsolver: Linear Mixed Model Solver* [R package version 1.0.2].

Brichet, N., Fournier, C., Turc, O., Strauss, O., Artzet, S., Pradal, C., Welcker, C., Tardieu, F., & Cabrera-Bosquet, L. (2017). A robot-assisted imaging pipeline for tracking the growths of maize ear and silks in a high-throughput phenotyping platform. *Plant Methods*, *13*(1), 1–12. https://doi.org/10.1186/s13007-017-0246-7

Brien, C. J., Berger, B., Rabie, H., & Tester, M. (2013). Accounting for variation in designing greenhouse experiments with special reference to greenhouses containing plants on conveyor systems. *Plant Methods*, *9*(1), 1–22. https://doi.org/10.1186/1746-4811-9-5

Brien, C. J., Jewell, N., Watts-Williams, S. J., Garnett, T., & Berger, B. (2020). Smoothing and extraction of traits in the growth analysis of noninvasive phenotypic data. *Plant Methods*, *16*(1), 1–21. https://doi.org/10.1186/s13007-020-00577-6

Brumback, B. A. (2010). On the built-in restrictions in linear mixed models, with application to smoothing spline analysis of variance. *Communications in Statistics—Theory and Methods*, *39*(4), 579–591. https://doi.org/10.1080/03610920902755847

Brumback, B. A., & Rice, J. A. (1998). Smoothing spline models for the analysis of nested and crossed samples of curves. *Journal of the American Statistical Association*, *93*(443), 961–976. https://doi.org/10.1080/01621459.1998.10473755

Buntaran, H., Piepho, H.-P., Schmidt, P., Rydén, J., Halling, M., & Forkman, J. (2020). Cross-validation of stage-wise mixed-model analysis of swedish variety trials with winter wheat and spring barley. *Crop Science*. https://doi.org/10.1002/csc2.20177

Bustos-Korts, D., Boer, M. P., Malosetti, M., Chapman, S., Chenu, K., Zheng, B., & van Eeuwijk, F. A. (2019). Combining crop growth modeling and statistical genetic modeling to evaluate phenotyping strategies. *Frontiers in Plant Science*, *10*. https://doi.org/10.3389/fpls.2019.01491

Butler, D. G., Cullis, B. R., Gilmour, A. R., Gogel, B. J., & Thompson, R. (2018). *ASReml-R reference manual Version 4*. VSN International Ltd, Hemel Hempstead, HP1 1ES, UK.

Cabrera-Bosquet, L., Fournier, C., Brichet, N., Welcker, C., Suard, B., & Tardieu, F. (2016). High-throughput estimation of incident light, light interception and radiation-use efficiency of thousands of plants in a phenotyping platform. *New Phytologist*, *212*(1), 269–281. https://doi.org/10.1111/nph.14027

Cui, E., Li, R., Crainiceanu, C. M., & Xiao, L. (2022). Fast multilevel functional principal component analysis. *Journal of Computational and Graphical Statistics*, 1–12. https://doi.org/10.1080/10618600.2022.2115500

Cullis, B. R., & Gleeson, A. C. (1991). Spatial analysis of field experiments - An extension to two dimensions. *Biometrics*, 1449–1460. https://doi.org/10.2307/2532398

Currie, I. D. (2014). Smooth mixed models for balanced longitudinal data. *Proceedings of the 29th International Workshop on Statistical Modelling, Goettingen, Germany*.

Currie, I. D., & Durban, M. (2002). Flexible smoothing with P-splines: A unified approach. *Statistical Modelling*, *2*(4), 333–349. https://doi.org/10.1191/1471082x02st039ob

Currie, I. D., Durban, M., & Eilers, P. H. C. (2006). Generalized linear array models with applications to multidimensional smoothing. *Journal of the Royal Statistical Society: Series B*, *68*(2), 259–280. https://doi.org/10.1111/j.1467-9868.2006.00543.x

Damesa, T. M., Möhring, J., Worku, M., & Piepho, H.-P. (2017). One step at a time: Stage-wise analysis of a series of experiments. *Agronomy Journal*, *109*(3), 845–857. https://doi.org/10.2134/agronj2016.07.0395

de Boor, C. (1978). *A Practical Guide to Splines* (Vol. 27). springer-verlag New York.

de Oliveira, H., Silva, F., Brito, L., Guarini, A., Jamrozik, J., & Schenkel, F. (2018). Comparing deregression methods for genomic prediction of test-day traits in dairy cattle. *Journal of Animal Breeding and Genetics*, *135*(2), 97–106. https://doi.org/10.1111/jbg.12317

Di, C.-Z., Crainiceanu, C. M., Caffo, B. S., & Punjabi, N. M. (2009). Multilevel functional principal component analysis. *The Annals of Applied Statistics*, *3*(1), 458. https://doi.org/10.1214/08-AOAS206SUPP

Djeundje, V. A. B., & Currie, I. D. (2010). Appropriate covariance-specification via penalties for penalized splines in mixed models for longitudinal data. *Electronic Journal of Statistics*, *4*, 1202–1224. https://doi.org/10.1214/10-EJS583

Durban, M., Hackett, C. A., McNicol, J. W., Newton, A. C., Thomas, W. T. B., & Currie, I. D. (2003). The practical use of semiparametric models in field trials. *Journal of Agricultural, Biological, and Environmental Statistics*, *8*(1), 48. https://doi.org/10.1198/1085711031265

Durban, M., Harezlak, J., Wand, M. P., & Carroll, R. J. (2005). Simple fitting of subject-specific curves for longitudinal data. *Statistics in Medicine*, *24*(8), 1153–1167. https://doi.org/10.1002/sim.1991

Eilers, P. H. C., Currie, I. D., & Durban, M. (2006). Fast and compact smoothing on large multidimensional grids. *Computational statistics & data analysis*, *50*(1), 61–76.

Eilers, P. H. C., & Marx, B. D. (1996). Flexible smoothing with b-splines and penalties. *Statistical Science*, 89–102. https://doi.org/10.1214/ss/1038425655

Eilers, P. H. C., & Marx, B. D. (2003). Multivariate calibration with temperature interaction using two-dimensional penalized signal regression. *Chemometrics and Intelligence Laboratory Systems*, *66*, 159–174. https://doi.org/10.1016/S0169-7439(03)00029-7

Eilers, P. H. C., & Marx, B. D. (2021). *Practical Smoothing: The Joys of P-splines*. Cambridge University Press.

Furrer, R., Flury, R., & Gerber, F. (2022). *spam: Sparse matrix* [R package version 2.9-1]. https://CRAN.R-project.org/package=spam

Garrick, D. J., Taylor, J. F., & Fernando, R. L. (2009). Deregressing estimated breeding values and weighting information for genomic regression analyses. *Genetics Selection Evolution*, *41*, 1–8. https://doi.org/10.1186/1297-9686-41-55

Gilmour, A. R., Cullis, B. R., & Verbyla, A. P. (1997). Accounting for natural and extraneous variation in the analysis of field experiments. *Journal of Agricultural, Biological, and Environmental Statistics*, 269–293. https://doi.org/10.2307/1400446

Goldsmith, J., Scheipl, F., Huang, L., Wrobel, J., Di, C., Gellar, J., Harezlak, J., McLean, M. W., Swihart, B., Xiao, L., Crainiceanu, C., & Reiss, P. T. (2022). *Refund: Regression with functional data* [R package version 0.1-26]. https://CRAN.R-project.org/package=refund

Green, P., Jennison, C., & Seheult, A. (1985). Analysis of field experiments by least squares smoothing. *Journal of the Royal Statistical Society: Series B*, *47*(2), 299–315. https://doi.org/10.1111/j.2517-6161.1985.tb01358.x

Greven, S., & Scheipl, F. (2017). A general framework for functional regression modelling. *Statistical Modelling*, *17*(1-2), 1–35. https://doi.org/10.1177/1471082X16681317

Guttorp, P., & Gneiting, T. (2006). Studies in the history of probability and statistics XLIX on the Matérn correlation family. *Biometrika*, *93*(4), 989–995. https://doi.org/10.1093/biomet/93.4.989

Hartung, J., Wagener, J., Ruser, R., & Piepho, H.-P. (2019). Blocking and re-arrangement of pots in greenhouse experiments: Which approach is more effective? *Plant Methods*, *15*(1), 1–11. https://doi.org/10.1186/s13007-019-0527-4

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, *72*(358), 320–338. https://doi.org/10.2307/2286796

Henderson, C. R. (1963). Selection index and expected genetic advance. *Statistical Genetics and Plant Breeding*, *982*, 141–163.

Hernández, M. A., Lee, D.-J., Rodríguez-Álvarez, M. X., & Durban, M. (accepted, 2023). Derivative curve estimation in longitudinal studies using P-splines. *Statistical Modelling*.

Hurtado, P. X., Schnabel, S. K., Zaban, A., Veteläinen, M., Virtanen, E., Eilers, P. H. C., van Eeuwijk, F. A., Visser, R. G. F., & Maliepaard, C. (2012). Dynamics of senescence-related QTLs in potato. *Euphytica*, *183*(3), 289–302. https://doi.org/10.1007/s10681-011-0464-4

Jin, X., Zarco-Tejada, P. J., Schmidhalter, U., Reynolds, M. P., Hawkesford, M. J., Varshney, R. K., Yang, T., Nie, C., Li, Z., Ming, B., et al. (2020). High-throughput estimation of crop traits: A review of ground and aerial phenotyping platforms. *IEEE Geoscience and Remote Sensing Magazine*, *9*(1), 200–231. https://doi.org/10.1109/MGRS.2020.2998816

Kar, S., Garin, V., Kholová, J., Vadez, V., Durbha, S. S., Tanaka, R., Iwata, H., Urban, M. O., & Adinarayana, J. (2020). Spatemhtp: A data analysis pipeline for efficient processing and utilization of temporal high-throughput phenotyping data. *Frontiers in Plant Science*, *11*, 1746. https://doi.org/10.3389/fpls.2020.552509

Kirchgessner, N., Liebisch, F., Yu, K., Pfeifer, J., Friedli, M., Hund, A., & Walter, A. (2016). The eth field phenotyping platform fip: A cable-suspended multi-sensor system. *Functional Plant Biology*, *44*, 154–168. https://doi.org/10.1071/FP16165

Kollers, S., Rodemann, B., Ling, J., Korzun, V., Ebmeyer, E., Argillier, O., Hinze, M., Plieske, J., Kulosa, D., Ganal, M. W., et al. (2013). Whole genome association mapping of fusarium head blight resistance in european winter wheat (triticum aestivum l.) *PLoS One*, *8*(2), e57500. https://doi.org/10.1371/journal.pone.0057500

Kronenberg, L., Yates, S., Boer, M. P., Kirchgessner, N., Walter, A., & Hund, A. (2021). Temperature response of wheat affects final height and the timing of stem elongation under field conditions. *Journal of Experimental Botany*, *72*(2), 700–717. https://doi.org/10.1093/jxb/eraa471

Kronenberg, L., Yu, K., Walter, A., & Hund, A. (2017). Monitoring the dynamics of wheat stem elongation: Genotypes differ at critical stages. *Euphytica*, *213*(7), 157. https://doi.org/10.1007/s10681-017-1940-2

Lee, D.-J. (2010). *Smoothing mixed model for spatial and spatio-temporal data* (Doctoral dissertation). Department of Statistics, Universidad Carlos III de Madrid.

Lee, D.-J., & Durban, M. (2011). P-spline anova-type interaction models for spatio-temporal smoothing. *Statistical Modelling*, *11*(1), 49–69. https://doi.org/10.1177/1471082X1001100104

Lee, D.-J., Durban, M., & Eilers, P. H. C. (2013). Efficient two-dimensional smoothing with p-spline anova mixed models and nested bases. *Computational Statistics & Data Analysis*, *61*, 22–37. https://doi.org/10.1016/j.csda.2012.11.013

Li, D., Quan, C., Song, Z., Li, X., Yu, G., Li, C., & Muhammad, A. (2021). High-throughput plant phenotyping platform (ht3p) as a novel tool for estimating agronomic traits from the lab to the field. *Frontiers in Bioengineering and Biotechnology*, *8*, 623705. https://doi.org/10.3389/fbioe.2020.623705

Li, Z., & Sillanpää, M. J. (2015). Dynamic quantitative trait locus analysis of plant phenomic data. *Trends in Plant Science*, *20*(12), 822–833. https://doi.org/10.1016/j.tplants.2015.08.012

Mead, R. (1997). Design of plant breeding trials. *Statistical methods for plant variety evaluation* (pp. 40–67). Springer. https://doi.org/10.1007/978-94-009-1503-9_4

Miao, C., Xu, Y., Liu, S., Schnable, P. S., & Schnable, J. C. (2020). Increased power and accuracy of causal locus identification in time series genome-wide association in sorghum. *Plant Physiology*, *183*(4), 1898–1909. https://doi.org/10.1104/pp.20.00277

Millet, E. J., Rodríguez-Álvarez, M. X., Perez-Valencia, D. M., Sanchez, I., Hilgert, N., van Rossum, B.-J., van Eeuwijk, F. A., & Boer, M. P. (2022). *statgenHTP: High Throughput Phenotyping (HTP) Data Analysis* [R package version 1.0.6]. https://CRAN.R-project.org/package=statgenHTP

Momen, M., Campbell, M. T., Walia, H., & Morota, G. (2019). Predicting longitudinal traits derived from high-throughput phenomics in contrasting environments using genomic legendre polynomials and b-splines. *G3: Genes, Genomes, Genetics*, *9*(10), 3369–3380. https://doi.org/10.1534/g3.119.400346

Montesinos-López, A., Montesinos-López, O. A., de los Campos, G., J, J. C., Burgueño, J., & Luna-Vazquez, F. J. (2018). Bayesian functional regression as an alternative statistical analysis of high-throughput phenotyping data of modern agriculture. *Plant Methods*, *14*, 1–17. https://doi.org/10.1186/s13007-018-0314-7

Montesinos-López, O. A., Montesinos-López, A., Crossa, J., de Los Campos, G., Alvarado, G., Suchismita, M., Rutkoski, J., González-Pérez, L., & Burgueño, J. (2017). Predicting grain yield using canopy

hyperspectral reflectance in wheat breeding data. *Plant Methods*, *13*, 1–23. https://doi.org/10.1186/s13007-016-0154-2

Moreira, F. F., Oliveira, H. R., Volenec, J. J., Rainey, K. M., & Brito, L. F. (2020). Integrating high-throughput phenotyping and statistical genomic methods to genetically improve longitudinal traits in crops. *Frontiers in Plant Science*, *11*, 681. https://doi.org/10.3389/fpls.2020.00681

Paine, C. E. T., Marthews, T. R., Vogt, D. R., Purves, D., Rees, M., Hector, A., & Turnbull, L. A. (2012). How to fit nonlinear plant growth models and calculate growth rates: An update for ecologists. *Methods in Ecology and Evolution*, *3*(2), 245–256. https://doi.org/10.1111/j.2041-210X.2011.00155.x

Patterson, H. D., & Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, *58*(3), 545–554. https://doi.org/10.1093/biomet/58.3.545

Perez-Valencia, D. M., Rodríguez-Álvarez, M. X., Boer, M. P., Kronenberg, L., Hund, A., Cabrera-Bosquet, L., Millet, E. J., & van Eeuwijk, F. A. (2022). A two-stage approach for the spatio-temporal analysis of high-throughput phenotyping data. *Scientific Reports*, *12*(1), 1–16. https://doi.org/10.1038/s41598-022-06935-9

Perez-Valencia, D. M., Rodríguez-Álvarez, M. X., Boer, M. P., & van Eeuwijk, F. A. (2023). A one-stage approach for the spatio-temporal analysis of high-throughput phenotyping data. *BioXiv*. https://doi.org/10.1101/2023.01.31.526411

Perich, G., Hund, A., Anderegg, J., Roth, L., Boer, M. P., Walter, A., Liebisch, F., & Aasen, H. (2020). Assessment of multi-image uav based high-throughput field phenotyping of canopy temperature. *Frontiers in Plant Science*, *11*, 150. https://doi.org/10.3389/fpls.2020.00150

Piepho, H.-P., Moehring, J., Schulz-Streeck, T., & Ogutu, J. O. (2012). A stage-wise approach for the analysis of multi-environment trials. *Biometrical Journal*, *54*(6), 844–860. https://doi.org/10.1002/bimj.201100219

Piepho, H.-P., Möhring, J., Melchinger, A. E., & Büchse, A. (2008). BLUP for phenotypic selection in plant breeding and variety testing. *Euphytica*, *161*(1), 209–228. https://doi.org/10.1007/s10681-007-9449-8

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

Piepho, H.-P., & Williams, E. R. (2010). Linear variance models for plant breeding trials. *Plant Breeding*, *129*(1), 1–8. https://doi.org/10.1111/j.1439-0523.2009.01654.x

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Core Team. (2019). *nlme: Linear and nonlinear mixed effects models* [R package version 3.1-142]. https://CRAN.R-project.org/package=nlme

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Ramsay, J. O., Graves, S., & Hooker, G. (2022). *fda: Functional Data Analysis* [R package version 6.0.5]. https://CRAN.R-project.org/package=fda

Ramsay, J. O., & Silverman, B. W. (2005). *Functional Data Analysis*. Springer New York.

Rodríguez-Álvarez, M. X., Boer, M. P., van Eeuwijk, F. A., & Eilers, P. H. C. (2018). Correcting for spatial heterogeneity in plant breeding experiments with p-splines. *Spatial Statistics*, *23*, 52–71. https://doi.org/10.1016/j.spasta.2017.10.003

Rodríguez-Álvarez, M. X., Durban, M., Lee, D.-J., & Eilers, P. H. C. (2019). On the estimation of variance parameters in non-standard generalised linear mixed models: Application to penalised smoothing. *Statistics and Computing*, (29), 483–500. https://doi.org/10.1007/s11222-018-9818-2

Rodríguez-Álvarez, M. X., Lee, D.-J., Kneib, T., Durban, M., & Eilers, P. H. C. (2015). Fast smoothing parameter separation in multidimensional generalized p-splines: The sap algorithm. *Statistics and Computing*, *25*(5), 941–957.

Roth, L., Rodríguez-Álvarez, M. X., van Eeuwijk, F. A., Piepho, H.-P., & Hund, A. (2021). Phenomics data processing: A plot-level model for repeated measurements to extract the timing of key stages and quantities at defined time points. *bioRxiv*. https://doi.org/10.1101/2021.05.02.442243

Ruppert, D., Wand, M. P., & Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge university press.

SAS Institute Inc. 2015. SAS/STAT®. (2015). *14.1 user's guide.* [Cary, NC: SAS Institute Inc].

Schmidt, P., Hartung, J., Bennewitz, J., & Piepho, H.-P. (2019). Heritability in plant breeding on a genotype-difference basis. *Genetics*, *212*(4), 991–1008. https://doi.org/10.1534/genetics.119.302134

Schulz-Streeck, T., Ogutu, J. O., & Piepho, H.-P. (2013). Comparisons of single-stage and two-stage approaches to genomic selection. *Theoretical and Applied Genetics*, *126*(1), 69–82. https://doi.org/10.1007/s00122-012-1960-1

Simpkin, A. J., Durban, M., Lawlor, D. A., MacDonald-Wallis, C., May, M. T., Metcalfe, C., & Tilling, K. (2018). Derivative estimation for longitudinal data analysis: Examining features of blood pressure measured repeatedly during pregnancy. *Statistics in Medicine*, *37*(19), 2836–2854. https://doi.org/10.1002/sim.7694

Slyusar, V. I. (1999). A family of face products of matrices and its properties. *Cybernetics and Systems Analysis*, *35*(3), 379–384.

Smith, A., Cullis, B., & Gilmour, A. (2001). Applications: The analysis of crop variety evaluation data in australia. *Australian & New Zealand Journal of Statistics*, *43*(2), 129–145. https://doi.org/10.1111/1467-842X.00163

Smith, S. P. (1995). Differentiation of the cholesky algorithm. *Journal of Computational and Graphical Statistics*, *4*(2), 134–147. https://doi.org/10.1080/10618600.1995.10474671

Song, P., Wang, J., Guo, X., Yang, W., & Zhao, C. (2021). High-throughput phenotyping: Breaking through the bottleneck in future crop breeding. *The Crop Journal*, *9*(3), 633–645. https://doi.org/10.1016/j.cj.2021.03.015

Tardieu, F., Cabrera-Bosquet, L., Pridmore, T., & Bennett, M. (2017). Plant phenomics, from sensors to knowledge. *Current Biology*, *27*(15), R770–R783. https://doi.org/10.1016/j.cub.2017.05.055

van Eeuwijk, F. A., Bustos-Korts, D., Millet, E. J., Boer, M. P., Kruijer, W., Thompson, A., Malosetti, M., Iwata, H., Quiroz, R., Kuppe, C., et al. (2019). Modelling strategies for assessing and increasing the effectiveness of new phenotyping techniques in plant breeding. *Plant Science*, *282*, 23–39. https://doi.org/10.1016/j.plantsci.2018.06.018

Velazco, J. G., Rodríguez-Álvarez, M. X., Boer, M. P., Jordan, D. R., Eilers, P. H. C., Malosetti, M., & van Eeuwijk, F. A. (2017). Modelling spatial trends in sorghum breeding field trials using a two-dimensional p-spline mixed model. *Theoretical and Applied Genetics*, *130*(7), 1375–1392. https://doi.org/10.1007/s00122-017-2894-4

Verbyla, A. P., Cullis, B. R., Kenward, M. G., & Welham, S. J. (1999). The analysis of designed experiments and longitudinal data by using smoothing splines. *Journal of the Royal Statistical Society: Series C*, *48*(3), 269–311. https://doi.org/10.1111/1467-9876.00154

Verbyla, A. P., De Faveri, J., Deery, D. M., & Rebetzke, G. J. (2021). Modelling temporal genetic and spatio-temporal residual effects for high-throughput phenotyping data. *Australian & New Zealand Journal of Statistics*. https://doi.org/10.1111/anzs.12336

Wand, M. P. (2003). Smoothing and mixed models. *Computational Statistics*, *18*(2), 223–249. https://doi.org/10.1007/s001800300142

Wang, R., Qiu, Y., Zhou, Y., Liang, Z., & Schnable, J. C. (2020). A high-throughput phenotyping pipeline for image processing and functional growth curve analysis. *Plant Phenomics*, *2020*. https://doi.org/10.34133/2020/7481687

Wang, R., Qiu, Y., Zhou, Y., Xu, Y., & Schnable, J. C. (2023). *implant: A High-throughput Phenotyping Pipeline for Image Processing and Functional Growth Curve Analysis* [R package version 0.1.0].

Welham, S., Cullis, B., Gogel, B., Gilmour, A., & Thompson, R. (2004). Prediction in linear mixed models. *Australian & New Zealand Journal of Statistics*, *46*(3), 325–347. https://doi.org/10.1111/j.1467-842X.2004.00334.x

Wolfinger, R. D. (1996). Heterogeneous variance: Covariance structures for repeated measures. *Journal of Agricultural, Biological, and Environmental Statistics*, 205–230. https://doi.org/10.2307/1400366

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R* (2nd ed.). Chapman; Hall/CRC.

Wood, S. N., & Scheipl, F. (2020). *gamm4: Generalized Additive Mixed Models using 'mgcv' and 'lme4'* [R package version 0.2-6]. https://CRAN.R-project.org/package=gamm4

Wood, S. N., Scheipl, F., & Faraway, J. J. (2013). Straightforward intermediate rank tensor product smoothing in mixed models. *Statistics and Computing*, *23*(3), 341–360.

Xiao, L., Zipunnikov, V., Ruppert, D., & Crainiceanu, C. (2016). Fast covariance estimation for high-dimensional functional data. *Statistics and Computing*, *26*, 409–421. https://doi.org/10.1007/s11222-014-9485-x

Xiao, Q., Bai, X., Zhang, C., & He, Y. (2022). Advanced high-throughput plant phenotyping techniques for genome-wide association studies: A review. *Journal of Advanced Research*, *35*, 215–230. https://doi.org/10.1016/j.jare.2021.05.002

Xu, Y., Li, Y., & Nettleton, D. (2018). Nested hierarchical functional data modeling and inference for the analysis of functional plant phenotypes. *Journal of the American Statistical Association*, *113*(522), 593–606. https://doi.org/10.1080/01621459.2017.1366907

Xu, Y., Li, Y., & Qiu, Y. (2021). Growth dynamics and heritability for plant high-throughput phenotyping studies using hierarchical functional data analysis. *Biometrical Journal*, *63*(6), 1325–1341. https://doi.org/10.1002/bimj.202000315

Xu, Y., Qiu, Y., & Schnable, J. C. (2018). Functional modeling of plant growth dynamics. *The Plant Phenome Journal*, *1*(1), 1–10. https://doi.org/10.2135/tppj2017.09.0007

Yang, W., Feng, H., Zhang, X., Zhang, J., Doonan, J. H., Batchelor, W. D., Xiong, L., & Yan, J. (2020). Crop phenomics and high-throughput phenotyping: Past decades, current challenges, and future perspectives. *Molecular Plant*, *13*(2), 187–214. https://doi.org/10.1016/j.molp.2020.01.008

Yang, W., Guo, Z., Huang, C., Duan, L., Chen, G., Jiang, N., Fang, W., Feng, H., Xie, W., Lian, X., et al. (2014). Combining high-throughput phenotyping and genome-wide association studies to reveal natural genetic variation in rice. *Nature Communications*, *5*(1), 5087.

Zhang, H. (2019). *Topics in functional data analysis and machine learning predictive inference* (Doctoral dissertation). Iowa State University.