# Spatial effect removal from field data by virtual replanting

Bachelor End Project Thesis


## Arnoud Glasbeek

Applied Mathematics



TU Delft

# Spatial effect removal from field data by virtual replanting

## Bachelor End Project Thesis

by

Arnoud Glasbeek

Student number: 5122473
Project duration: April 19, 2022 –July 15, 2022
Thesis committee: Dr. N. V. Budko, TU Delft, supervisor
Dr. D. J. P. Lahaye, TU Delft

TU Delft

# Abstract

In agricultural studies it is often important to predict the performance of genetically different plants. To make sure predictions are done well, it is necessary to make sure they are not influenced by effects of the field on which they are planted. These field effects or spatial effects are in practice often quite complicated and can be due to a wide variety of reasons. To get a better view of these field effects a good mathematical model is desired. In this paper a model is presented which helps to find these field effects. This model tries to estimate the field effect by comparing data of the same plant on different positions of the field. Data is obtained in a finite amount of positions, which means that the model finds the field effect in a finite amount of positions as well. This field effect is found using a cross-validation technique obtained from Tikhonov regularization. The field effect in a finite amount of positions is extended to a field effect in every position of the field. To do this in a good way a kernel method is used, the advantage of which is that it does not depend on a mesh. This kernel method is here applied with a kernel function that is based on Gaussian distribution. This model is applied to several fields of crops to get a view of the performance of the model on real data.

# Table of Contents

# Introduction

In agricultural fields it is common to have genetically different plants and plants with different production origin. Plants of the same genotype and production origin are planted at different, often random, locations in the test field to avoid the so-called field effect. The field effect can be caused by a wide variety of things, ranging from nutrients in the soil being different across the field to shadows causing some parts of the field to receive less sunlight. The exact causes of the field effect are not important for this research, which is mainly concerned with the removal of the field effect from the measured phenotype data and thus giving a better estimate of the performance of different plants.
There exists substantial research into the field effect and multiple models have been proposed to improve the crop/phenotype estimates. Examples of this can be found in **(Rodríguez-Álvarez et al., 2018)** and **(Salvador et al., 2022)**. These models usually try to estimate both a fixed and a random effect on the field. Using these effects a model is build that is able to estimate the expected phenotype from the data at different locations.
The approach in this project is slightly different. The fields that are observed in this project have plants of different production origin. The field is divided into small parts (*plots*), with each plot containing the same amount of plants. Thus, for each production-origin group of plants, called *batch*, multiple plots exist at different positions of the field.
By comparing the mean phenotype of the plants from the same batch, the field effect can be estimated, as this should be the only thing causing the differences in mean phenotype for these plants, apart from the statistical error, which is expected to be small. Using this estimated field effect all data can be virtually moved to one position of the field. After this virtual movement or replanting, the results for different batches can be compared better. Because of the checking how plants do in another position this model is called the Replanting Model.
In this project the reconstruction of the field effect is performed with a kernel method. This method is usually used in machine learning to find relations between large sets of data **(Hofmann et al., 2008)**.
The advantage of the kernel method is that it can represent and reconstruct a function of the field effect without the need to construct a mesh on the basis of plot centroids. Such meshes often turn out to be highly irregular and affect the quality of results.
This report starts of by fully explaining how the model works in Chapter 1. In Chapter 2 the kernel method is introduced. In Chapter 3 the structure of the data is explained in detail. Along with this one of the parameters, that is needed to find the field effect and heavily relies on the data, is found. Chapter 4 is devoted to finding another important parameter of the model. In Chapter 5 the results of the kernel method on some fields are presented.

# Chapter 1

# Replanting model

In this project we will be looking at a model which we call the Replanting Model. The idea of this model is to transform all obtained data to one point on the field and compare the yield that would be observed there. This process of transforming the data will here be called virtual replanting. Because of this virtual replanting all field conditions should become the same, which means that the spatial field effect would be removed.
To model the fields, the fields are divided into small plots of land. On each plot several plants are planted, divided over ridges. Every plant has a position on the field, which is denoted by $\mathbf{r} = (x, y)^T$, for some $x$ and $y$ within the boundaries of the field. To give a plot on which several plants are planted a position the average of these positions is taken. This average position is called a centroid and this is the point in which data about the plot is obtained. The amount of plots on the field is denoted by $P$.
The plants planted on the field are genetically different, which can influence the yield of a plant quite strongly. The plants are of several varieties, and every variety is divided into batches. Crops within the same batch are assumed to be exactly the same and thus there is assumed to be no difference between plants of the same batch, other than statistical error. The crops are divided into $B$ batches and on each plot plants of the same batch are planted. This means that there are $\frac{P}{B} = K$ plots for each batch of plants. The plants are planted across the field in groups of the same race, resulting in large parts of the field being planted with the same race. For each variety there are multiple groups of plots and these groups are placed across the field in no particular order. In each of these groups one plot of each batch is planted. These plots are again placed in no particular order.
An example of the way in which the plots are planted can be seen in Figure 1.1. Here each variety of plants, in this case potatoes of different varieties, is given by a different colour. This will be done in this way for all the figures of the fields presented in this report. Each plot is marked with the number of the batch of the potato that is planted. The values on the axes denote the position of the plot.
To form a good model for this problem we will first look at a probabilistic model. Using this probabilistic model a statistical model is made, from which, using regularization, an equation is obtained that can be solved to obtain the function for the field effect.

## 1.1 Probabilistic model

Let $Y_b$ be a random variable that gives the yield of a batch $b$. This random variable has a conditional probability density function $\rho_b(y|\mathbf{r}_p)$. For this density function we assume a spatial transformation. Using this spatial transformation the distribution of batch $b$ at location $\mathbf{r}_p$ can be obtained in terms of the distribution at another location $\mathbf{r}_q$. This transformation is given by

$$\rho_b(y|\mathbf{r}_p) = \phi_1 \rho(\phi_0 + \phi_1 y|\mathbf{r}_q), \tag{1.1}$$

where $\phi_0$ and $\phi_1$ are functions of $\mathbf{r}_p$ and $\mathbf{r}_q$, with $\phi_1(\mathbf{r}_p, \mathbf{r}_q) > 0$.
For these functions it is assumed that the multiplicative part, $\phi_1$, is equal to 1, which means that only the additive transformation remains. This gives

$$\rho_b(y|\mathbf{r}_p) = \rho_b(\phi_0 + y|\mathbf{r}_q). \tag{1.2}$$

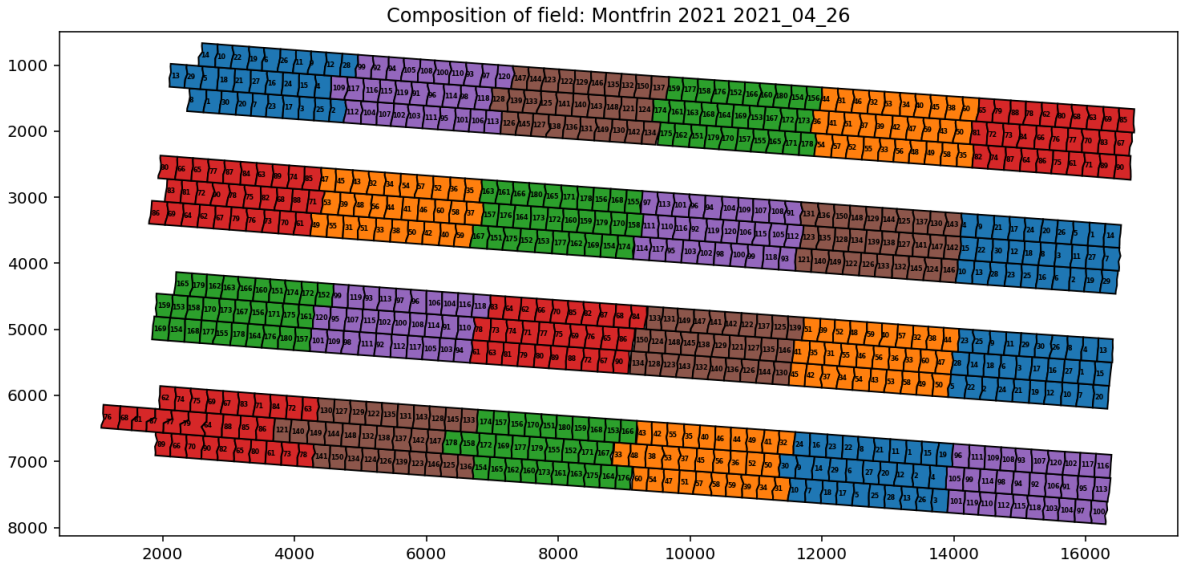



Figure 1.1: An example of a field with different varieties of crops. Each color gives one race, the numbers denote the different batches.

This probability distribution can be used to express the expectations and variance of the yield in different locations. For example the expectation of $Y_b$ at $\mathbf{r}_q$ is obtained from

$$\mathbb{E}[Y_b|\mathbf{r}_q] = \int y\rho_b(y|\mathbf{r}_q)\,dy \tag{1.3}$$

Now we can express the expectation and variance in another position, $\mathbf{r}_q$, in terms of this.

$$\begin{aligned}\mathbb{E}[Y_b|\mathbf{r}_p] &= \int y\rho_b(y|\mathbf{r}_p)\,dy \\ &= \int y\rho_b(\phi_0 + y|\mathbf{r}_q)\,dy \\ &= \int (-\phi_0 + y')\rho_b(y|\mathbf{r}_q)\,d(-\phi_0 + y') \\ &= -\phi_0 + \mathbb{E}[Y_b|\mathbf{r}_q]\end{aligned} \tag{1.4}$$

And for variance

$$\begin{aligned}\mathrm{Var}\left[Y_b|\mathbf{r}_p\right] &= \int \left(y - \mathbb{E}[Y_b|\mathbf{r}_p]\right)^2 \rho_b(y|\mathbf{r}_p)\,dy \\ &= \int \left(-\phi_0 + y' + \phi_0 - \mathbb{E}\left[Y_b|\mathbf{r}_q\right]\right)^2 \rho_b(y'|\mathbf{r}_q)\,d\left(-\phi_0 + y'\right) \\ &= \int \left(y' - \mathbb{E}\left[Y_b|\mathbf{r}_q\right]\right)^2 \rho_b(y'|\mathbf{r}_q)\,dy' = \mathrm{Var}[Y_b|\mathbf{r}_q].\end{aligned} \tag{1.5}$$

Now Equation 1.4 can be written out for both $\mathbb{E}[Y_b|\mathbf{r}_p]$ and $\mathbb{E}[Y_b|\mathbf{r}_q]$. Along with what is obtained in Equation 1.5 the following is obtained:

$$\begin{aligned}\mathbb{E}[Y_b|\mathbf{r}_p] &= -\phi_0(\mathbf{r}_p, \mathbf{r}_q) + \mathbb{E}[Y_b|\mathbf{r}_q] \\ \mathbb{E}[Y_b|\mathbf{r}_q] &= -\phi_0(\mathbf{r}_q, \mathbf{r}_p) + \mathbb{E}[Y_b|\mathbf{r}_p] \\ \mathrm{Var}\left[Y_b|\mathbf{r}_p\right] &= \mathrm{Var}\left[Y_b|\mathbf{r}_q\right]\end{aligned} \tag{1.6}$$

From this it is found that $\phi_0$ must be such that

$$-\phi_0(\mathbf{r}_p, \mathbf{r}_q) = \phi_0(\mathbf{r}_q, \mathbf{r}_q) \tag{1.7}$$

To satisfy this function, it can be chosen such that $\phi_0(\mathbf{r}_p, \mathbf{r}_q) = \gamma(\mathbf{r}_p) - \gamma(\mathbf{r}_q)$ for some function $\gamma$. Obviously the function can also be chosen in another way, but this representation is able to find the function on a finite

amount of locations. Since the data is always obtained on a finite number of locations this is good enough. With this Equations 1.4 and 1.5 can be rewritten in terms of $\gamma$

$$\begin{aligned} \gamma(\mathbf{r}_p) + \mathbb{E}[Y_b|\mathbf{r}_p] &= \gamma(\mathbf{r}_q) + \mathbb{E}[Y_b|\mathbf{r}_q] \\ \mathrm{Var}\left[Y_b|\mathbf{r}_p\right] &= \mathrm{Var}\left[Y_b|\mathbf{r}_q\right] \end{aligned} \tag{1.8}$$

To now accomplish the set goal of virtually replanting the data, this $\gamma$ function must be found.

## 1.2 Statistical model

Of each plant the size of the canopy is measured using aerial pictures. From these aerial pictures the size of the leaves of the plants can be calculated. The size of these leaves, also called the canopy, gives a good approximation for the size of the plant itself and thus for the yield when harvesting the plants. Using this canopy data for each plot the average, $E_b(\mathbf{r})$, and the standard deviation, $S_b(\mathbf{r})$ can be obtained. These are estimates of the actual expectation, $\mathbb{E}[Y_b|\mathbf{r}]$, and standard deviation $\mathrm{Std}[Y_b|\mathbf{r}]$. This average and standard deviation are obtained in one place on a plot, called the centroid of the plot. This centroid is gives the average positions of the plants planted on a plot. It is assumed that the statistical error between these estimates $E_b(\mathbf{r})$ and $S_b(\mathbf{r})$ and $\mathbb{E}[Y_b|\mathbf{r}]$ and $\mathrm{Std}[Y_b|\mathbf{r}]$ are approximately normally distributed. These errors can be denoted by $\varepsilon_E$ and $\varepsilon_S$. As the standard deviation will not be used any further in this model, $\varepsilon_S$ will not be used and $\varepsilon$ will be used for $\varepsilon_E$. In this research we do not go further into these errors, as we will be relying on a cross-validation technique to find the function $\gamma$, which will be explained later on.
Now as the data is only obtained in a finite number of points, the centroid points, the $\gamma$-function will also only be obtained in these same finite number of centroids. This is not what is desired, as this $\gamma$-function is desired to be found in any position on the field. To find $\gamma(\mathbf{r})$, we need to introduce so called basis functions, given by $V_p(\mathbf{r})$, which will help to get $\gamma(\mathbf{r})$ for any $\mathbf{r}$ on the field. With $\gamma_p$ as the value of $\gamma$ in the position of the centroid of plot $p$ and basis function $V_p(\mathbf{r})$ for the same plot, which is a function of the position in which $\gamma$ is desired, $\gamma(\mathbf{r})$ can be obtained, by summing over these terms for all plots:

$$\gamma(\mathbf{r}) = \sum_{p=1}^{P} \gamma_p V_p(\mathbf{r}). \tag{1.9}$$

The term $V_p(\mathbf{r})$ can be obtained in several ways, but is usually obtained using Finite Element Method basis functions **(Hughes et al., 2005)**. The problem with these functions is that they depend on a mesh of the field, which can be very bad. The way to make a mesh for a field like this is by forming triangles between two neighbouring plots in one row and one neighbouring plot from the next row. Because the plots of the fields often are not always placed nicely next to each other, these triangles are not always of the same form. An example of this is the first and second row in Figure 1.1. In these rows the distance between the centers of the first plots of the two rows is quite large. This results in triangles that have one side that is much larger than the other two. This gives a triangle that is much different from the triangles obtained between plots from other rows. A mesh like this, with large differences between the different partitions can influence the results in a bad way.
From the probabilistic relations established in Equation 1.8 and the experimental data we can obtain the linear system for one batch $b$

$$\gamma(\mathbf{r}_p^{(b)}) - \gamma(\mathbf{r}_q^{(b)}) + E_b(\mathbf{r}_p^{(b)}) - E_b(\mathbf{r}_q^{(b)}) = \varepsilon_b(r_p^{(b)}) - \varepsilon_b(r_q^{(b)}) \tag{1.10}$$

For each plot $p$ in this batch there can be K different equations of this form, comparing with each possible $q$ within the same batch. One of these equations obviously would not give any useful information, as it would have $\mathbf{r}_p$ twice, which would just give 0 on both sides. This leaves $K-1$ useful equations to determine $K$ variables, which gives an underdetermined system.
These $K-1$ equations can be written in several different ways and to do this one of the plots must be chosen to be compared with all the other plots. The positions of the plots can be numbered $\mathbf{r}_1^{(b)} \cdots \mathbf{r}_K^{(b)}$. Then the first plot, at position $\mathbf{r}_1^{(b)}$, is chosen as the plot with which the others are compared. How this numbering is done is not very important and does not influence the result. From this the following equations are obtained:

$$\begin{aligned} \gamma(\mathbf{r}_1^{(b)}) - \gamma(\mathbf{r}_2^{(b)}) + E_b(\mathbf{r}_1^{(b)}) - E_b(\mathbf{r}_2^{(b)}) &= \varepsilon_b(r_1^{(b)}) - \varepsilon_b(r_2^{(b)}) \\ &\vdots \\ \gamma(\mathbf{r}_1^{(b)}) - \gamma(\mathbf{r}_K^{(b)}) + E_b(\mathbf{r}_1^{(b)}) - E_b(\mathbf{r}_K^{(b)}) &= \varepsilon_b(r_1^{(b)}) - \varepsilon_b(r_K^{(b)}) \end{aligned} \tag{1.11}$$

This can be rewritten to a matrix equation:

$$\begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ 1 & 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \cdots & -1 \end{bmatrix} \begin{pmatrix} \gamma(\mathbf{r}_1^{(b)}) \\ \gamma(\mathbf{r}_2^{(b)}) \\ \vdots \\ \gamma(\mathbf{r}_K^{(b)}) \end{pmatrix} + \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ 1 & 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \cdots & -1 \end{bmatrix} \begin{pmatrix} E_b(\mathbf{r}_1^{(b)}) \\ E_b(\mathbf{r}_2^{(b)}) \\ \vdots \\ E_b(\mathbf{r}_K^{(b)}) \end{pmatrix} = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ 1 & 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \cdots & -1 \end{bmatrix} \begin{pmatrix} \varepsilon_b(\mathbf{r}_1^{(b)}) \\ \varepsilon_b(\mathbf{r}_2^{(b)}) \\ \vdots \\ \varepsilon_b(\mathbf{r}_K^{(b)}) \end{pmatrix} \tag{1.12}$$

This results in an equation without the fully written matrices and vectors that looks like this

$$R_b\gamma_b + R_b\mathbf{E}_b = R_b\varepsilon_b \tag{1.13}$$

Where the variables assigned here denote the matrices and vectors from the equation above.
These vectors and matrices can be expanded upon to form a system for all batches together. To do this we let $R$ be the Kronecker product of an identity matrix $I_B$ and the $R_b$ matrices for all batches:

$$R = I_B \otimes R_b. \tag{1.14}$$

This gives a non-square matrix $R$: $R \in \mathrm{R}^{(P-B)\times P}$. Vectors $\gamma$, $\mathbf{E}$ and $\varepsilon$ are made by combining vectors $\gamma_b$, $\mathbf{E}_b$ and $\varepsilon_b$ for all batches. These vectors are all vectors of length $P$. This gives a matrix-vector system for all data:

$$R\gamma + R\mathbf{E} = R\varepsilon \tag{1.15}$$

## 1.3 Regularization

The system obtained in the previous section is, as was already seen, underdetermined. This is because of the $R$ matrix being a non-square matrix. Because of this and the presence of the statistical error $\varepsilon$ minimizing $R\gamma + R\mathbf{E}$ will lead to over-fitting. To avoid over-fitting a method is obtained from Tikhonov regularization **(van Wieringen, 2015)**. Here a parameter $\eta$ is introduced, which is the regularization parameter. Using this a way to find $\gamma$ is found.

$$\begin{aligned} \gamma &= \arg\min_{\eta} F_\eta(\gamma) \\ F_\eta(\gamma) &= \frac{1}{2}||R\gamma + R\mathbf{E}||_2^2 + \frac{\eta}{2}\int_\Omega |\nabla\gamma(r)|^2\, d\Omega \end{aligned} \tag{1.16}$$

The vector $\gamma$ that satisfies this equation can be found by solving the Euler-Lagrange equation

$$(R^T R + \eta D)\gamma_b = -R^T R\mathbf{E}, \tag{1.17}$$

where the matrix $D$ is introduced. This Euler-Lagrange equation is derived in Appendix A The matrix $D$ in this system is a stiffness matrix, which can be obtained in several ways. The most common way to do this is using the Finite Element Method **(Reddy, 2006)**.
For this parameter $\eta$ still a optimal value must be found, which is done using cross-validation. How this is done exactly will be further explained in Chapter 4.

# Chapter 2

# Kernel Method with the Gaussian Kernel

In the previous chapter we saw that there were some terms in the equations which were obtained from applying the Finite Element Method, like the term $V_p(\mathbf{r})$ in Equation 1.9 and the matrix $D$ in Equation 1.17. The problem with FEM is that there will always be effects of the mesh visible in the $\gamma$-function, especially on fields where it is hard to make a good mesh, like the fields used in this project. This is not desired, as the approximation of the $\gamma$ function is wanted to be as good as possible. Thus it is desired to find a method which performs better.
A proposed method for this is the so called kernel method, which is often used in machine learning **(Hofmann et al., 2008)**. For this a function $\kappa(\mathbf{r_1}, \mathbf{r_2})$ is introduced, where the $\mathbf{r_1}$ and $\mathbf{r_2}$ are two positions on a field. Using these the term $V_p(\mathbf{r})$ in the $\gamma$-function, introduced in 1.9, can be given by $V_p(\mathbf{r}) = \kappa(\mathbf{r}, \mathbf{r_p})$. These new $V_p$ are not influenced by a mesh, which means that they can give a more accurate approximation of the field effect.
The kernel function is usually chosen to be a radial basis function, so that

$$\kappa(\mathbf{r_1}, \mathbf{r_2}) = \kappa(|\mathbf{r_1} - \mathbf{r_2}|) \tag{2.1}$$

where $|\mathbf{r_1} - \mathbf{r_2}|$ gives the vector norm or distance, given by $|\mathbf{r}| = \sqrt{r_x^2 + r_y^2}$.
The addition of the kernel function transforms the system 1.12 into the following system

$$RK\gamma + R\mathbf{E} = -R\varepsilon \tag{2.2}$$

Here the matrix $K$ is given by the kernel $\kappa$, by

$$[K]_{p,q} = \kappa(\mathbf{r}_p, \mathbf{r}_q), \tag{2.3}$$

where $\mathbf{r}_p$ and $\mathbf{r}_q$ are the positions plots which correspond to the entry of $K$ at position $[p, q]$.
This changes the function to be minimized to

$$F_\eta(\gamma) = \frac{1}{2}||RK\gamma + R\mathbf{E}||_2^2 + \frac{\eta}{2}\int_{\mathbb{R}^2} |\nabla\gamma_h(\mathbf{r})|^2 \, d\mathbf{r} \tag{2.4}$$

With Euler-Lagrange equation

$$(K^T R^T RK + \eta D)\gamma = -K^T R^T R\mathbf{E} \tag{2.5}$$

Where matrix $D$ is given by

$$[D]_{p,q} = \int_{\mathbb{R}^2} \nabla\kappa(\mathbf{r}, \mathbf{r_p}) \cdot \nabla\kappa(\mathbf{r}, \mathbf{r_q}) \, d\mathbf{r} \tag{2.6}$$

From this equation $\gamma$ can be found at the positions of the centroids, at which data is obtained. Then Equation 1.9, with now terms $V_p$ obtained from the kernel function, can be used to find $\gamma$ in any position.
This kernel function can be chosen in various ways. In this research we will focus on one kernel function, the Gaussian kernel. This proposed kernel, based on a Gaussian distribution, is given by

$$\kappa(\mathbf{r_1}, \mathbf{r_2}, \sigma) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{|\mathbf{r}_1 - \mathbf{r}_2|^2}{2\sigma^2}\right), \tag{2.7}$$

with $\sigma > 0$. This gives $K$-matrix

$$[K]_{p,q} = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{|\mathbf{r}_p - \mathbf{r}_q|^2}{2\sigma^2}\right) \tag{2.8}$$

## 2.1 Stiffness matrix

The stiffness matrix $D$ for the Gaussian kernel can be found by calculating its elements using the Equation 2.6. For the Gaussian kernel we have:

$$\nabla\kappa(\mathbf{r_1},\mathbf{r_2},\sigma) = -\frac{\mathbf{r_1}-\mathbf{r_2}}{2\pi\sigma^4}\exp\left(-\frac{|\mathbf{r}_1-\mathbf{r}_2|^2}{2\sigma^2}\right). \tag{2.9}$$

Putting this in Equation 2.6 gives the following terms for the D-matrix

$$D_{p,q} = \frac{1}{4\pi^2\sigma^8}\int_{\mathrm{R}^2}(\mathbf{r}-\mathbf{r_p})\cdot(\mathbf{r}-\mathbf{r_q})\exp\left(-\frac{1}{2\sigma^2}\left(|\mathbf{r}-\mathbf{r}_p|^2+|\mathbf{r}-\mathbf{r}_q|^2\right)\right)d\mathbf{r} \tag{2.10}$$

Now we set $(\mathbf{r}-\mathbf{r_p}) = \mathbf{r}'$. Then this can be worked out in the following way

$$\begin{aligned} D_{p,q} &= \frac{1}{4\pi^2\sigma^8}\int_{\mathrm{R}^2}\mathbf{r}'\cdot(\mathbf{r}'+\mathbf{r_p}-\mathbf{r_q})\exp\left(-\frac{1}{2\sigma^2}\left(|\mathbf{r}'|^2+|\mathbf{r}'+\mathbf{r_p}-\mathbf{r}_q|^2\right)\right)d\mathbf{r} \\ &= \frac{1}{4\pi^2\sigma^8}\int_{\mathrm{R}^2}|\mathbf{r}'|^2\exp\left(-\frac{1}{2\sigma^2}\left(2|\mathbf{r}'|^2+|\mathbf{r}_p-\mathbf{r}_q|^2+2\mathbf{r}'\cdot(\mathbf{r_p}-\mathbf{r_q})\right)\right) \\ &+\mathbf{r}'\cdot(\mathbf{r_p}-\mathbf{r_q})\exp\left(-\frac{1}{2\sigma^2}\left(2|\mathbf{r}'|^2+|\mathbf{r}_p-\mathbf{r}_q|^2+2\mathbf{r}'\cdot(\mathbf{r_p}-\mathbf{r_q})\right)\right)d\mathbf{r} \end{aligned} \tag{2.11}$$

Now to make the terms look simpler, $|\mathbf{r}'| = r$ and $|\mathbf{r}_p-\mathbf{r}_q|$. When doing this one can obtain, from the dot product, that $\mathbf{r}'\cdot(\mathbf{r_p}-\mathbf{r_q}) = rs\cos(\theta)$, where $\theta$ is the angle between the two vectors. With this the equation for $D$ can be rewritten as

$$\begin{aligned} [D]_{p,q} = \frac{1}{4\pi^2\sigma^8}\int_0^\infty\int_0^{2\pi} & r^3\exp\left(-\frac{1}{2\sigma^2}\left(2r^2+s^2+2rs\cos(\theta)\right)\right) \\ & + r^2 s\cos(\theta)\exp\left(-\frac{1}{2\sigma^2}\left(2r^2+s^2+2rs\cos(\theta)\right)\right)d\theta\,dr \end{aligned} \tag{2.12}$$

In this integral the modified Bessel function can be recognised **(Arfken et al., 2013)**, with

$$\begin{aligned} \int_0^{2\pi}\exp\left(-\frac{rs}{\sigma^2}\cos\theta\right)d\theta &= 2\pi I_0\left(\frac{rs}{\sigma^2}\right) \\ \int_0^{2\pi}\cos(\theta)\exp\left(-\frac{rs}{\sigma^2}\cos\theta\right)d\theta &= -2\pi I_1\left(\frac{rs}{\sigma^2}\right) \end{aligned} \tag{2.13}$$

This gives

$$\begin{aligned} [D]_{p,q} = \frac{1}{2\pi\sigma^8}\exp\left(-\frac{s^2}{2\sigma^2}\right)\int_0^\infty & r^3 I_0\left(\frac{rs}{\sigma^2}\right)\exp\left(-\frac{r^2}{\sigma^2}\right) \\ & - sr^2 I_1\left(\frac{rs}{\sigma^2}\right)\exp\left(-\frac{r^2}{\sigma^2}\right)dr \end{aligned} \tag{2.14}$$

Here $I_0$ and $I_1$ give the Bessel functions of the first kind, with series representation

$$\begin{aligned} I_0\left(\frac{rs}{\sigma^2}\right) &= \sum_{n=0}^{\infty}\left(\frac{s}{\sigma^2}\right)^{2n}\frac{1}{4^n n!n!}r^{2n} \\ I_1\left(\frac{rs}{\sigma^2}\right) &= \left.\frac{dI_0(z)}{dz}\right|_{z=rs/\sigma^2} = \sum_{n=1}^{\infty}\left(\frac{s}{\sigma^2}\right)^{2n-1}\frac{2n}{4^n n!n!}r^{2n-1} \end{aligned} \tag{2.15}$$

Now we can rewrite 2.14 in a way such that is two integrals, one over the $I_0$ function and one over the $I_1$ function:

$$\begin{aligned} [D]_{p,q} = \frac{1}{2\pi\sigma^8}\exp\left(-\frac{s^2}{2\sigma^2}\right)\Bigg(\int_0^\infty & r^3 I_0\left(\frac{rs}{\sigma^2}\right)\exp\left(-\frac{r^2}{\sigma^2}\right)dr \\ & -\int_0^\infty sr^2 I_1\left(\frac{rs}{\sigma^2}\right)\exp\left(-\frac{r^2}{\sigma^2}\right)dr\Bigg) \end{aligned} \tag{2.16}$$

We will elaborate these two integrals separately. We will start of with the integral for the $I_0$ Bessel's function, where we will substitute the series expansion for the Bessel's equation

$$\int_0^\infty r^3 I_0\left(\frac{rs}{\sigma^2}\right)\exp\left(-\frac{r^2}{\sigma^2}\right)dr = \int_0^\infty r^3\sum_{n=0}^{\infty}\left(\frac{s}{\sigma^2}\right)^{2n}\frac{1}{4^n n!n!}r^{2n}\exp\left(-\frac{r^2}{\sigma^2}\right)dr \tag{2.17}$$

Here we can interchange the summation and integration, since for $r \in [0,\infty)$, positive $\sigma$ and $s$, which we both assume to be positive, this is allowed. This gives us the series

$$\sum_{n=0}^{\infty} \left(\frac{s}{\sigma^2}\right)^{2n} \frac{1}{4^n n!n!} \int_0^{\infty} r^{2n+3} \exp\left(-\frac{r^2}{\sigma^2}\right) dr \tag{2.18}$$

The integral within this series is known and can be evaluated giving

$$\sum_{n=0}^{\infty} -\left(\frac{s}{\sigma^2}\right)^{2n} \frac{1}{4^n n!n!} \frac{1}{2\sigma^{-2n-4}} \left[\Gamma(n+2, -\frac{r^2}{\sigma^2})\right]_0^{\infty} \tag{2.19}$$

Here $\Gamma(a,x)$ gives the incomplete Gamma function, which can be evaluated at 0 and $\infty$, with $\Gamma(n+2,0) = (n+1)!$ and $\Gamma(n+2,\infty) = 0$. Putting this in 2.19 gives

$$\begin{aligned} -\sum_{n=0}^{\infty} \left(\frac{s}{\sigma^2}\right)^{2n} \frac{1}{4^n n!n!} \frac{1}{2\sigma^{-2n-4}} \left[\Gamma(n+2, -\frac{r^2}{\sigma^2})\right]_0^{\infty} &= \sum_{n=0}^{\infty} \left(\frac{s}{\sigma^2}\right)^{2n} \frac{1}{4^n n!n!} \frac{1}{2\sigma^{-2n-4}} (n+1)! \\ &= \sum_{n=0}^{\infty} \left(\frac{s}{\sigma^2}\right)^{2n} \frac{(n+1)!}{2 \cdot 4^n n!n! \sigma^{-2n-4}} \\ &= \sum_{n=0}^{\infty} \frac{s^{2n}}{2\sigma^{2n-4}} \cdot \frac{n+1}{4^n n!} \\ &= \sum_{n=0}^{\infty} \frac{n+1}{2 \cdot 4^n n! \sigma^{2n-4}} s^{2n} \end{aligned} \tag{2.20}$$

This last sum is a known series, with

$$\sum_{n=0}^{\infty} \frac{n+1}{2 \cdot 4^n n! \sigma^{2n-4}} s^{2n} = \frac{1}{8} \exp\left(\frac{s^2}{4\sigma^2}\right) \sigma^2 (4\sigma^2 + s^2) \tag{2.21}$$

So now we have the first integral, we will do the same for the second part.

$$\int_0^{\infty} sr^2 I_1\left(\frac{rs}{\sigma^2}\right) \exp\left(-\frac{r^2}{\sigma^2}\right) dr = \int_0^{\infty} sr^2 \sum_{n=1}^{\infty} \left(\frac{s}{\sigma^2}\right)^{2n-1} \frac{2n}{4^n n!n!} r^{2n-1} \exp\left(-\frac{r^2}{\sigma^2}\right) dr \tag{2.22}$$

Again summation and integration can be interchanged to obtain

$$\sum_{n=1}^{\infty} \left(\frac{s}{\sigma^2}\right)^{2n-1} \frac{2n}{4^n n!n!} \int_0^{\infty} r^{2n+1} \exp\left(-\frac{r^2}{\sigma^2}\right) dr \tag{2.23}$$

Now we can evaluate this integral in the same way as in Equation 2.19, to get

$$\sum_{n=1}^{\infty} -\left(\frac{s}{\sigma^2}\right)^{2n-1} \frac{n}{4^n n!n!} \frac{1}{\sigma^{-2n-2}} \left[\Gamma(n+1, -\frac{r^2}{\sigma^2})\right]_0^{\infty} \tag{2.24}$$

Now we can do the same with this as we have done in Equation 2.20

$$\begin{aligned} \sum_{n=1}^{\infty} -\left(\frac{s}{\sigma^2}\right)^{2n-1} \frac{n}{4^n n!n!} \frac{1}{\sigma^{-2n-2}} \left[\Gamma(n+1, -\frac{r^2}{\sigma^2})\right]_0^{\infty} &= \sum_{n=1}^{\infty} \left(\frac{s}{\sigma^2}\right)^{2n-1} \frac{n}{4^n n!n!} \frac{1}{2\sigma^{-2n-2}} n! \\ &= \sum_{n=1}^{\infty} \left(\frac{s}{\sigma^2}\right)^{2n-1} \frac{nn!}{4^n n!n! \sigma^{-2n-4}} \\ &= \sum_{n=1}^{\infty} \left(\frac{s}{\sigma^2}\right)^{2n-1} \frac{n}{4^n n! \sigma^{-2n-4}} \\ &= \sum_{n=1}^{\infty} \frac{n}{\sigma^{2n-4} 4^n n!} s^{2n-1} \end{aligned} \tag{2.25}$$

Again this sum is known, with

$$\sum_{n=1}^{\infty} \frac{n}{\sigma^{2n-2} 4^n n!} s^{2n-1} = \frac{1}{4} \exp\left(\frac{s^2}{4\sigma^2}\right) s \tag{2.26}$$

Now we can put this back in Equation 2.16, to obtain

$$\begin{aligned}
[D]_{p,q} &= \frac{1}{2\pi\sigma^8}\exp\left(-\frac{s^2}{2\sigma^2}\right)\left(\frac{1}{8}\exp\left(\frac{s^2}{4\sigma^2}\right)\sigma^2(4\sigma^2+s^2)-\frac{1}{4}\exp\left(\frac{s^2}{4\sigma^2}\right)s\right)\\
&= \frac{1}{8\pi\sigma^8}\exp\left(-\frac{s^2}{2\sigma^2}\right)\exp\left(\frac{s^2}{4\sigma^2}\right)\left(\frac{1}{2}\sigma^2(4\sigma^2+s^2)-s\right)\\
&= \frac{1}{8\pi\sigma^8}\exp\left(-\frac{s^2}{2\sigma^2}+\frac{s^2}{4\sigma^2}\right)\left(\frac{1}{2}\sigma^2(4\sigma^2+s^2)-s\right)\\
&= \frac{1}{8\pi\sigma^8}\exp\left(-\frac{s^2}{4\sigma^2}\right)\left(\frac{1}{2}\sigma^2 s^2-s+2\sigma^4\right)\\
&= \frac{1}{8\pi\sigma^8}\exp\left(-\frac{|\mathbf{r}_p-\mathbf{r}_q|^2}{4\sigma^2}\right)\left(\frac{1}{2}\sigma^2|\mathbf{r}_p-\mathbf{r}_q|^2-|\mathbf{r}_p-\mathbf{r}_q|+2\sigma^4\right)
\end{aligned} \tag{2.27}$$

## 2.2 Properties of the $K$ and $D$ matrices

In Section 2.1 two different matrices were obtained, the $K$- and $D$- matrices, given by

$$\begin{aligned}
[K]_{p,q} &= \frac{1}{2\pi\sigma^2}\exp\left(-\frac{|\mathbf{r}_p-\mathbf{r}_q|^2}{2\sigma^2}\right)\\
[D]_{p,q} &= \frac{1}{8\pi\sigma^8}\exp\left(-\frac{|\mathbf{r}_p-\mathbf{r}_q|^2}{4\sigma^2}\right)\left(\frac{1}{2}\sigma^2|\mathbf{r}_p-\mathbf{r}_q|^2-|\mathbf{r}_p-\mathbf{r}_q|+2\sigma^4\right).
\end{aligned} \tag{2.28}$$

It is desireable to know more about the properties of these matrices.
First of all it is the case that these matrices are square $P\times P$ matrices, where $P$ is the number of plots. Secondly it is easily visible that both matrices are symmetric, since $|\mathbf{r}_p-\mathbf{r}_q| = |\mathbf{r}_q-\mathbf{r}_p|$, and thus $[K]_{p,q} = [K]_{q,p}$ and $[D]_{p,q} = [D]_{q,p}$. Since $|\mathbf{r}_p-\mathbf{r}_p| = 0$, the values on the diagonal of both matrices can be obtained, with $[K]_{p,p} = \frac{1}{2\pi\sigma^2}$ and $[D]_{p,p} = \frac{1}{4\pi\sigma^4}$. It is also certain that both matrices have only real and positive entries. For $K$ this is always the case and for $D$ this follows from the way in which $\sigma$ was chosen in the previous section. Since both matrices only have real values these matrices are Hermitian matrices.

### 2.2.1 Positive entries of $D$ -matrix

To make the model work well, it is desired to have all of the entries of the matrices be non-negative. For $K$ this is obviously always the case, as both $\sigma$ and $|\mathbf{r}_p-\mathbf{r}_q|$ are positive. For $D$ this is less obvious. The entries of $D$ are, if for any $s = |\mathbf{r}_p-\mathbf{r}_q|$ and $\sigma$,

$$\frac{1}{8\pi\sigma^8}\exp\left(-\frac{s^2}{4\sigma^2}\right)\left(\frac{1}{2}\sigma^2 s^2-s+2\sigma^4\right)\geq 0 \tag{2.29}$$

So if

$$\frac{1}{2}\sigma^2 s^2-s+2\sigma^4\geq 0, \tag{2.30}$$

since $\frac{1}{8\pi\sigma^8}\exp\left(-\frac{s^2}{4\sigma^2}\right) > 0$.
Equation 2.30 can be solved with equality, and then we obtain two zeroes

$$\begin{aligned}
s_1 &= \frac{2-\sqrt{4-16\sigma^6}}{2\sigma^2}\\
s_2 &= \frac{2+\sqrt{4-16\sigma^6}}{2\sigma^2}
\end{aligned} \tag{2.31}$$

Now it is desired to know for which $s$ and which $\sigma$ Equation 2.30 does hold. It is immediately clear that if $4-16\sigma^6 < 0$, the roots of this equation are complex, which means that $\frac{1}{2}\sigma^2 s^2-s+2\sigma^4 > 0$ for any real $s$. This follows from the fact that $\sigma > 0$, and thus $2\sigma^4 > 0$. So for $s = 0$ this equation is possible. Since the function has no real root and is positive in one point, it is positive everywhere. Thus Equation 2.30 holds for all real $s$ if $4-16\sigma^6 < 0$, or $\sigma^2 > \frac{1}{\sqrt[3]{4}}$.

Next it is desired to know when Equation 2.30 holds for $\sigma^2 \leq \frac{1}{\sqrt[3]{4}}$. It is known that between the two zeroes found, $s_1$ and $s_2$ in Equation 2.31, the polynomial in Equation 2.30 is negative, so if $s > s_2$ or $s < s_1$ for all $s$ then Equation 2.30 does hold. Now $\sigma$ can be chosen in a way such that, for all $s$ which are obtained with the data that is being used, $s > s_2$. It turns out that this is possible, but with one exception, $s = 0$ (so when $\mathbf{r}_p = \mathbf{r}_q$), which is always present in the matrix $D$. Now it is observed that Equation 2.30 holds for $s = 0$, since $\sigma$ is positive. This means that this exception does not cause trouble, as the desired property will still be satisfied.
Now since it is that $\sigma > 0$,

$$\begin{aligned} s_2 &< \frac{2+\sqrt{4}}{2\sigma^2} \\ s_2 &< \frac{4}{2\sigma^2} \\ s_2 &< \frac{2}{\sigma^2}. \end{aligned} \tag{2.32}$$

If we now let $m = \min_{p,q}(|\mathbf{r}_p - \mathbf{r}_q|)$, so the minimal distance between the centroids of two plots on the field and thus the minimal value $s$ can have with the obtained data, then we desire to have that $m > s_2$. Now $\sigma$ can be chosen such that this is the case.
Here $m > s_2$ if $m \geq \frac{2}{\sigma^2}$, so if $\sigma^2 \geq \frac{2}{m}$ or $\sigma \geq \sqrt{\frac{2}{m}}$.
This gives us the desired bounds for $\sigma^2$ and $\sigma$

$$\begin{aligned} \sigma^2 &\geq \frac{2}{m} \\ \sigma &\geq \sqrt{\frac{2}{m}} \end{aligned} \tag{2.33}$$

No upper bound for $\sigma$ exists for this property, as it was obtained earlier that Equation 2.30 is always positive for large values of $\sigma$.

### 2.2.2 Positive definiteness of the $K$- and $D$ -matrix

Something else that is desirable to know whether the matrices are positive definite. This is because these types of matrices have some computational advantages. To show that they are positive definite it suffices to show that they are strictly diagonally dominant, which is the case for some matrix $A$

$$|A_{pp}| > \sum_{i \neq p} |A_{pi}|, \tag{2.34}$$

So if the entry on the diagonal is larger than the sum of the entries of the rest of the row, for each row, then the matrix is strictly diagonally dominant. A strictly diagonally dominant matrix with only real and positive entries is always positive definite.
It will be shown that, using a bound for $\sigma$, both $K$ and $D$ can satisfy this property. This will be shown in the following two lemmas.

**Lemma 2.2.1.** *If $\sigma^2 < \frac{m^2}{\ln(P-1)}$, where $m = \min_{p,q}(|\mathbf{r}_p - \mathbf{r}_q|)$ and $P$ is the number of plots, then matrix $K$ is strictly diagonally dominant, and since $K$ only has positive real entries, it is positive definite.*

*Proof.* Suppose $\sigma^2 < \frac{m^2}{2\ln P-1}$, then

$$\begin{aligned} \ln(n-1) &< \frac{m^2}{2\sigma^2} \\ n-1 &< \exp(\frac{m^2}{\sigma^2}) \\ \frac{1}{n-1} &> \frac{1}{\exp(\frac{m^2}{\sigma^2})} \\ 1 &> (n-1)\exp(-\frac{m^2}{2\sigma^2}) \end{aligned} \tag{2.35}$$

Now here it can be noted that $\exp\left(-\frac{|\mathbf{r}_p-\mathbf{r}_i|^2}{2\sigma^2}\right) < \exp\left(-\frac{m^2}{2\sigma^2}\right)$, since $|\mathbf{r}_p - \mathbf{r}_i| > m$ and the negative exponential decreases if its argument becomes larger. Thus $(n-1)\exp\left(-\frac{m^2}{2\sigma^2}\right) > \sum_{i\neq p}\exp\left(-\frac{|\mathbf{r}_p-\mathbf{r}_i|^2}{2\sigma^2}\right)$, for any $p$, where $p$

is the index of one plot on the field and $i$ is the index of another plot. Thus this is a sum over the distances between one plot, $p$, and all other plots.
From this it is obtained that

$$1 > \sum_{i \neq p} \exp\left(-\frac{|\mathbf{r}_p - \mathbf{r}_i|^2}{2\sigma^2}\right) \tag{2.36}$$

Multiplying both sides with $\frac{1}{2\pi\sigma^2}$ now gives

$$\frac{1}{2\pi\sigma^2} > \sum_{i \neq p} \frac{1}{2\pi\sigma^2} \exp\left(-\frac{|\mathbf{r}_p - \mathbf{r}_i|^2}{2\sigma^2}\right) \tag{2.37}$$

Here the left hand side denotes the diagonal values of matrix $K$ and the right hand side denotes the sum over the non-diagonal values in some row, which corresponds to the plot with index $p$. As this holds for any $p$, and thus for any row of $K$, it is obtained that $K$ is strictly diagonally dominant. □

The same can be done for matrix $D$, although it is a bit more complicated, by proving the following lemma

**Lemma 2.2.2.** *Let $m = \min_{p,q}(|\mathbf{r}_p - \mathbf{r}_q|)$ Let $x$ be such that $\frac{\sqrt[3]{16(P-1)}}{\ln(P-1)} < x < \frac{\ln(P-1)}{m}$ and $P < \exp(\frac{m}{x}) + 1$. Then let $y$ be such that*

$$y > \ln\left(\frac{(x\ln(P-1))^3}{16} - P + 2\right) + 1 \tag{2.38}$$

*Then if $\frac{4}{m} < \sigma^2 < \frac{x}{4y}m$, matrix $D$ is strictly diagonally dominant, and since $D$ only has positive real entries, it is positive definite.*

*Proof.* It is assumed that

$$y > \ln\left(\frac{(x\ln(P-1))^3}{16} - P + 2\right) + 1 \tag{2.39}$$

Which is the same as

$$y > \frac{\ln\left(\frac{(x\ln(P-1))^3}{16} + 1\right)}{\ln(P-1)} + 1 \tag{2.40}$$

This can be rewritten to

$$\begin{aligned} y - 1 > \frac{\ln\left(\frac{(x\ln(P-1))^3}{16} + 1\right)}{\ln(P-1)} \quad (P-1)^{y-1} &> \frac{(x\ln(P-1))^3}{16} + 1 \\ 1 &> \frac{1}{(P-1)^{y-1}}\left(\frac{(x\ln(P-1))^3}{16} + 1\right) \\ 1 &> \frac{P-1}{\exp(\ln(P-1))^y}\left(\frac{(x\ln(P-1))^3}{16} + 1\right) \\ 1 &> (P-1)\exp\left(-y\ln(P-1)\right)\left(\frac{(x\ln(P-1))^3}{16} + 1\right). \end{aligned} \tag{2.41}$$

Now since $P < \exp\frac{m}{x} + 1$, from which it is obtained that $m > 6\ln(P-1)$. With this and the fact that the right hand side of the equation decreases as the argument $6\ln(P-1)$ increases, for $P > 20$ Equation 2.41 can be rewritten as

$$1 > (P-1)\exp\left(-\frac{y}{x}m\right)\left(\frac{m^3}{16} + 1\right). \tag{2.42}$$

From the conditions on $\sigma^2$ two things follow. Firstly, from $\sigma^2 > \frac{4}{m}$, it follows that $\left(\frac{m^3}{16} + 1\right) > \left(\frac{m^2}{4\sigma^2} - \frac{m}{2\sigma^4} + 1\right)$. Secondly, from $\sigma^2 < \frac{x}{4y}m$, it follows that $\exp\left(-\frac{y}{x}m\right) > \exp\left(-\frac{m^2}{4\sigma^2}\right)$. Using this Equation 2.42 can be rewritten as

$$1 > (P-1)\exp\left(-\frac{m^2}{4\sigma^2}\right)\left(\frac{m^2}{4\sigma^2} - \frac{m}{2\sigma^4} + 1\right). \tag{2.43}$$

Here a term $\frac{1}{2\sigma^4}$ can be taken in front, which results in

$$1 > \frac{P-1}{2\sigma^4}\exp\left(-\frac{m^2}{4\sigma^2}\right)\left(\frac{1}{2}\sigma^2 m^2 - \frac{m}{\sigma^2} + 1\right). \tag{2.44}$$

Now, since $m = \min_{p,q}(|\mathbf{r}_p - \mathbf{r}_q|)$,

$$\frac{P-1}{2\sigma^4}\exp\left(-\frac{m^2}{4\sigma^2}\right)\left(\frac{1}{2}\sigma^2 m^2 - \frac{m}{\sigma^2} + 1\right) > \sum_{i\neq p}\frac{1}{2\pi\sigma^4}\exp\left(-\frac{|\mathbf{r}_p - \mathbf{r}_i|^2}{4\sigma^2}\right)\left(\frac{1}{2}\sigma^2|\mathbf{r}_p - \mathbf{r}_i|^2 - |\mathbf{r}_p - \mathbf{r}_i| + 2\sigma^4\right). \tag{2.45}$$

The right hand side of the inequality in Equation 2.43 is decreasing if $m$ becomes larger for $m > 6\ln(n-1)$, so from this and Equation 2.44 it can be obtained that

$$1 > \sum_{i\neq p}\frac{1}{2\pi\sigma^4}\exp\left(-\frac{|\mathbf{r}_p - \mathbf{r}_i|^2}{4\sigma^2}\right)\left(\frac{1}{2}\sigma^2|\mathbf{r}_p - \mathbf{r}_i|^2 - |\mathbf{r}_p - \mathbf{r}_i| + 2\sigma^4\right). \tag{2.46}$$

Now both sides can be multiplied with $\frac{1}{4\pi\sigma^4}$, to obtain

$$\frac{1}{4\pi\sigma^4} > \sum_{i\neq p}\frac{1}{8\pi\sigma^8}\exp\left(-\frac{|\mathbf{r}_p - \mathbf{r}_i|^2}{4\sigma^2}\right)\left(\frac{1}{2}\sigma^2|\mathbf{r}_p - \mathbf{r}_i|^2 - |\mathbf{r}_p - \mathbf{r}_i| + 2\sigma^4\right) \tag{2.47}$$

Here it is observed that the left hand side is exactly the entry on the diagonal of $D$ and the right hand side is the sum over all non-diagonal entries in the same row. Because this inequality holds for the assumed bounds on $\sigma$, $m$ and $P$, matrix $D$ is strictly diagonally dominant and thus positive definite. □

The bounds in this lemma seem to be more complicated than they actually are. Both $P$ and $m$ can be obtained immediately from the data at hand, thus an upper bound of $x$ can be found quite easily. Using that it can be observed that $y$ can be obtained, and that the upper bound for $\sigma$ is as large as possible if $y$ is as small as possible. Thus $y$ can always be chosen such that it is a small as possible to satisfy Equation 2.38, which can be obtained from an equality in this equation. From this the upper bound for $\sigma$ can be obtained by maximizing the function obtained from $\frac{x}{4y}$. This maximum is, nearly always, attained at the largest value for $x$.
Now it has to be noted that the bounds for $\sigma$ found using these lemmas are not necessarily the smallest bounds for $\sigma$, but these bounds are very useful for finding actual bounds. When $m$ and $P$ are obtained from data on the field actual bounds for positive definiteness can be found using iteration. This is done in Chapter 3.

# Chapter 3

# Composition of data

## 3.1 Obtained data

To test this model data on crops is needed. This data is obtained from three potato fields, two in The Netherlands and one in France. The data consists of canopy data for each ridge on the field, obtained on several days in multiple years for each field. Along with the canopy data data is obtained for the positions of the ridges, which means that the fields can be reconstructed. For each field in each year one of the days is chosen to be further observed.
The fields from which the data is obtained are located in Montfrin in France and SPNA and Veenklooster, both in The Netherlands. These fields are observed in three different years, 2019, 2020 and 2021, resulting in nine different datasets. Each field consists of plots of four ridges, with 6 crops in each ridge. On each field six different varieties of potatoes are planted. Exact properties of these varieties are not important, only that all are different types of potatoes. Each variety is planted on one sixth of the plots on the field.
On each of the fields the same batches are planted on the same amount of plots. There are 180 batches planted on 4 plots each. This mean that on each field 720 plots are planted. This gives the parameters which were used in Chapter 1 to be:

- Amount of plots $P = 720$
- Amount of batches $B = 180$
- Plots per batch $K = 4$

Since For each of the fields the data is such that $K = 4$, it is possible to construct the blocks of which the matrix $R$ consists. This gives

$$R_b = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 1 & 0 & 0 & -1 \end{bmatrix}. \tag{3.1}$$

Something else that can be observed immediately on the field is the minimal distance between the centroids, which was already used in Chapter 2 and denoted by $m$. The crops are planted roughly the same each year. This should mean that the coordinates of the plots and distances between them are the same for different years as well, but unfortunately this is not the case. Sometimes crops are planted slightly different across the years, which means that the coordinates vary between the years. Along with this the data is scaled differently in different years. It is possible to rescale the data such that the fields are similar for each year, but even then there will always be differences between the years. Because of this it is chosen to not scale the fields, and use the original data. The only problem that occurs with this is that then for each year the bounds for $\sigma$ are different and thus have to be found separately.
In Figures 3.1, 3.2 and 3.3 the three different fields can be seen in the way they were planted in 2021.
In Figure 3.4 the canopy data of the field in Montfrin in 2021 can be seen. Here a dark green color means a low yield and a light green color means a high yield. On fields like these it is desired to find spatial effect and to correct the data, such that the data is similar across the field. Similar data is available for all the fields, and can be found in Appendix B.

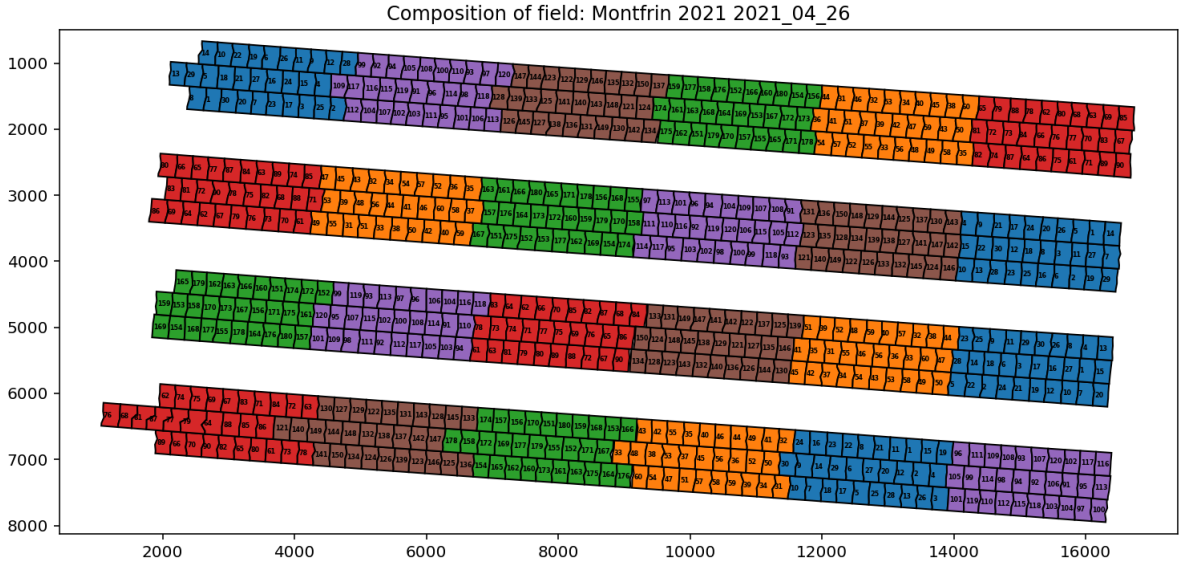



Figure 3.1: Field composition of the field in Montfrin in 2021

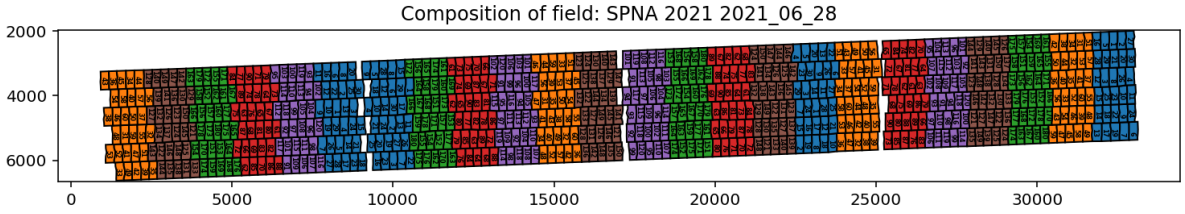



Figure 3.2: Field composition of the field in SPNA in 2021

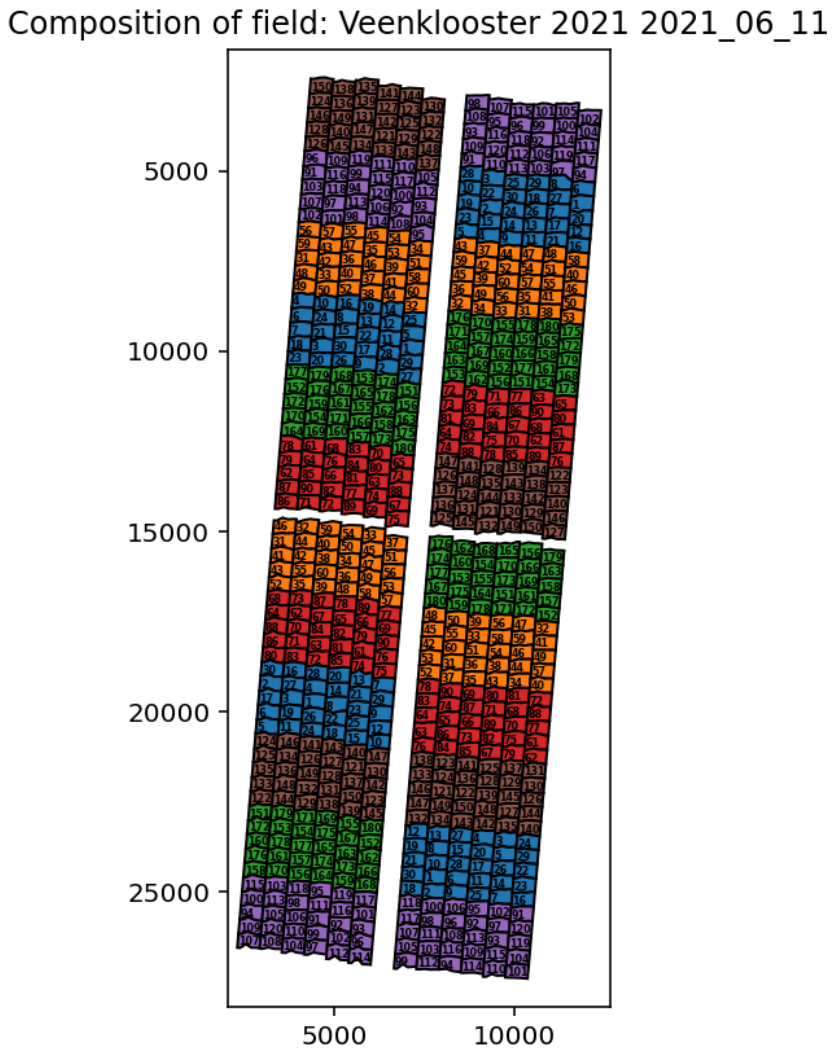



Figure 3.3: Field composition of the field in Veenklooster in 2021

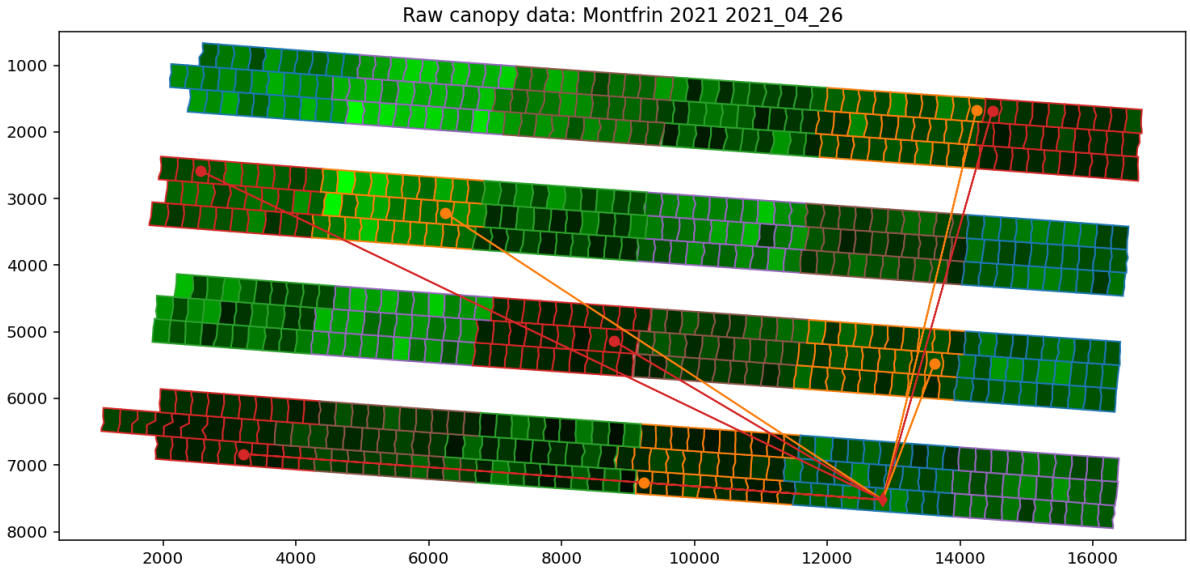



Figure 3.4: Raw data of the field in Montfrin in 2021. A dark shade denotes a low yield, a light color a high yield.

## 3.2 Finding the $\sigma$ parameter

### 3.2.1 Bounds for $\sigma$

In Chapter 2 methods were established for finding an upper and lower bound for $\sigma$. These bounds depend on the distance between plots and the amounts of plots on the field. The amount of plots is always the same, but the distance between plots differs across the different fields and across the years. This means that nine different bounds for $\sigma$ must be obtained. Here the calculation, using the lemmas in Chapter 2, for these bounds will be done for the field in Montfrin in the year 2021. The bounds obtained by doing this calculation for the other fields can be seen in the fourth (lower bound) and fifth (upper bound) columns of Table 3.1.
Now in the Lemmas 2.2.1 and 2.2.2 some restrictions were made to make sure the lemmas work for any type of matrix. This means that the bounds obtained from these lemmas are not necessarily the largest and smallest values of $\sigma$ for which the matrices are positive definite. It is not even the case that the largest value for which the matrices are diagonally dominant, which is used in the lemmas, is also the largest value for which the matrices are positive definite. Now the most important is the upper bound of $\sigma$, since the values for $\sigma$ which will be used must be large. The reason for this is that the kernel function, introduced in Equation 2.3, goes to zero quickly if $\sigma$ is small and $|\mathbf{r}_1 - \mathbf{r}_2|$ increases. This means that the term $V_p$ introduced in Chapter 1 and given for the Kernel Method by

$$V_p(\mathbf{r}) = \kappa(\mathbf{r}, \mathbf{r}_p, \sigma) \tag{3.2}$$

will become to small if $\sigma$ is small. If this term is very small the field effect will be observed as zero in any point where no data is observed. If this is the case the entire purpose of this method is lost, since this method is used to find the field effect in points in which no data is known.
By iterating over increasingly large $\sigma$ an estimation for the true upper bound can be found. This iteration is started with $\sigma$ as the value of upper bound found from calculation and increases until a $\sigma$ is found for which one of the matrices is not positive definite anymore. For each $\sigma$ the matrices $K$ and $D$ are computed and the eigenvalues of these matrices are calculated. Then the positive definiteness of these is checked by checking whether the smallest eigenvalues (and thus all eigenvalues) are positive. If the matrices have no negative eigenvalues, they are positive definite, since both are symmetric matrices **(Horn and Johnson, 1985)**. However, if one of the eigenvalues is negative, the matrix is not positive definite. The upper bound obtained from this is usually found to be about 5 to 10 times as large as the upper bound which is obtained earlier. In the sixth column these actual upper bounds for positive definiteness can be seen for the different fields.

### 3.2.2 Choice of $\sigma$

These bounds are used to choose a fixed value for $\sigma$ with which the field effect is estimated. This $\sigma$ must be fixed in a way in which it gives as much information about the field effect as possible and that the information that is obtained is accurate.
First suppose that $\sigma$ is chosen to be small. If $\sigma$ is chosen to be small, it can be obtained that the Gaussian kernel $\kappa(\mathbf{r}, \mathbf{r}_p, \sigma)$ is very small for $\mathbf{r}$ and $\mathbf{r}_p$, except of course if $\mathbf{r} = \mathbf{r}_p$. Now this means that $V_p(\mathbf{r})$ will be small for any $\mathbf{r}$ that is not $\mathbf{r}_p$. The field effect $\gamma(\mathbf{r})$ was found from the sum over the product of $V_p(\mathbf{r})$ and the field effect in the centroid positions, $\gamma(\mathbf{r}_p)$,

$$\gamma(\mathbf{r}) = \sum_{p=1}^{P} \gamma_p V_p(\mathbf{r}). \tag{3.3}$$

Since $V_p(\mathbf{r})$ is very small if $\mathbf{r} \neq \mathbf{r}_p$, this field effect is very small everywhere, except in the centroid locations. This means that the only information that is obtained from the field effect is the field effect in the centroid locations. This information was already known, as these where given by the $\gamma_p$, which means that this method does not add any new information if $\sigma$ is chosen to be small.
Secondly suppose $\sigma$ is chosen to be as large as possible, so very close to the upper bound found before. Then it is obtained that the information obtained on the field effect will not be very accurate. What this information and the accuracy of the information exactly is will be further explained in Section 4.2. To make sure that $\sigma$ is chosen to be not too large and not too small, $\sigma$ is chosen at around half of the upper bound obtained before. In Chapter 4 it is obtained that this choice gives enough information about the field effect and that the obtained information is quite accurate. Thus the values for $\sigma$ that are chosen are half of the upper bound found by iteration, rounded to the nearest integer. The values that are chosen for all the fields can be seen in the seventh

column of Table 3.1.
For these values of $\sigma$ the matrices $K$ and $D$ both look very close to diagonal matrices. The non-diagonal elements are often more than a factor 100 smaller than the diagonal elements. Even when choosing $\sigma$ closer to the upper bounds found by iteration the matrices will look close to diagonal. The only elements that are then somewhat of the same order as the values on the diagonal are the values on the first of diagonal. Although this might seem strange this is often what is seen in stiffness matrices for other methods as well.

| Field | Year | Minimal Distance ($m$) | Lower bound | Upper bound for $\sigma$ from calculation | Upper bound for $\sigma$ from iteration | $\sigma$ used |
|---|---|---|---|---|---|---|
| Montfrin | 2019 | 48.2 | 0.082 | 2.64 | 19.4 | 10 |
| | 2020 | 234.4 | 0.131 | 11 | 96.7 | 48 |
| | 2021 | 227.7 | 0.133 | 10.7 | 90.1 | 45 |
| SPNA | 2019 | 96.7 | 0.20 | 4.91 | 38.6 | 19 |
| | 2020 | 251.75 | 0.126 | 11.7 | 100.9 | 50 |
| | 2021 | 245 | 0.128 | 11.42 | 99.6 | 50 |
| Veenklooster | 2019 | 184.9 | 0.147 | 8.83 | 73 | 37 |
| | 2020 | 239.6 | 0.129 | 11.19 | 100 | 50 |
| | 2021 | 378.4 | 0.102 | 17.0 | 149.8 | 75 |

Table 3.1: Bounds for $\sigma$ on the different fields

### 3.2.3 Calculation on Montfrin 2021

Here the calculation from which the values in the table for Montfrin 2021 will be done more extensively. For the Montfrin field in 2021, $m \approx 227.1$ and $P = 720$.
From this the lower bound for $\sigma$ can be obtained fairly easily. In the lemmas in Chapter 2 it was found that $\sigma$ must be such that $\sigma^2 \geq \frac{4}{m}$ or $\sigma \geq \sqrt{\frac{4}{m}}$. This means that for this field $\sigma \geq \sqrt{\frac{4}{227.1}} \approx 0.133$. Now since $\sigma$ is chosen to be large, this bound is easily satisfied.
In the lemmas in Chapter 2 it can be seen that Lemma 2.2.2 is far more restrictive than Lemma 2.2.1. This means that if a upper bound is found that satisfies Lemma 2.2.2, it will nearly always satisfy Lemma 2.2.1 as well. Thus we will first look at the lemma with the most restrictive conditions, Lemma 2.2.2, and then check whether the other lemma is satisfied as well.
In Lemma 2.2.2 it is obtained that $D$ is diagonally dominant if $\sigma < \frac{x}{4y} m$, where $x$ and $y$ are such that $\frac{\sqrt[3]{16(P-1)}}{\ln(P-1)} < x < \frac{\ln(P-1)}{m}$ and

$$y > \ln\left(\frac{(x\ln(P-1))^3}{16} - P + 2\right) + 1 \tag{3.4}$$

These can be found using the values for $m$ and $P$, $m = 227.1$ and $P = 720$. This means that $x$ must be such that $\frac{\sqrt[3]{16 \cdot 719}}{\ln(719)} < x < \frac{227.1}{\ln(719)}$, which can be rounded to $3.43 < x < 34.5$. Now we must find $y$ such that it satisfies

$$y > \ln\left(\frac{(x\ln(719))^3}{16} - 718\right) + 1 \tag{3.5}$$

and such that the upper bound for $\frac{x}{4y} m$, is maximal. Thus a maximum for $\frac{x}{4y}$ must be found, since $m$ is constant. Now from the way in which $x$ is chosen we know for the following for the term inside the logarithm:

$$\begin{aligned} x &> \frac{\sqrt[3]{16 \cdot 719}}{\ln(719)} \\ x\ln(719) &> \sqrt[3]{16 \cdot 719} \\ \frac{(x\ln(719))^3}{16} &> 719 \\ \frac{(x\ln(719))^3}{16} - 718 &> 1. \end{aligned} \tag{3.6}$$

From this it follows that the logarithm in the bound for $y$ is positive, and thus $y$ has a lower bound of at least 1. From this lower bound and the fact that $\frac{x}{4y}$ decreases if $y$ increases for some fixed $x$ if $y > 1$, it follows that the maximum is always attained at the smallest possible value for $y$ for a given $x$. Thus the maximum of $\frac{x}{4y}$ is the maximum of

$$\begin{aligned} g(x) &= \frac{x}{4\left(\ln\left(\frac{(x\ln(719))^3}{16} - 718\right) + 1\right)} \\ g(x) &= \frac{x}{4\ln\left((x\ln(719)^3 - 718\right) + 4}. \end{aligned} \tag{3.7}$$

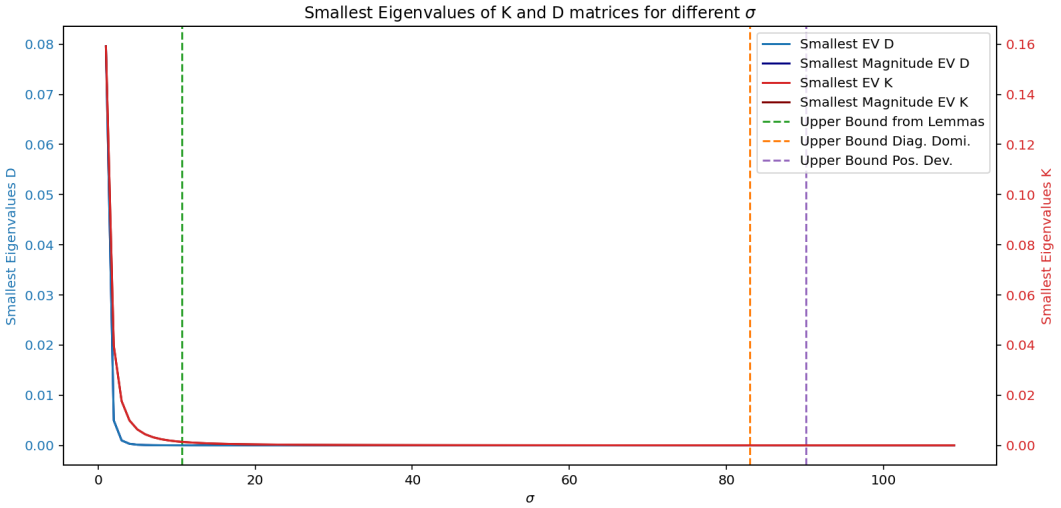
It can be observed that $g(x)$ is an increasing function, so the maximum is attained at the maximal value of x, here given by $\frac{m}{\ln(P-1)} \approx 34.5$. This gives a upper bound for $\frac{x}{4y}$ at $g(34.5) \approx 0.499$.
Thus we can use bounds for $\sigma$ as $\frac{4}{m} \leq \sigma^2 < 0.499m$. For the $m$ obtained from the data this gives $0.0176 \leq \sigma^2 < 113.4$, or $0.133 \leq \sigma < 10.7$. Now it is easy to check that this value satisfies lemma 2.2.1 as well. For this lemma $\sigma$ must be such that $\sigma^2 < \frac{m^2}{\ln(P-1)}$. Here we can see that this would mean that $\sigma^2 < 7838.85$ or $\sigma < 88.53$. This means that the upper bound found lemma 2.2.2 satisfy lemma 2.2.1 as well.
Now we can try iteration over $\sigma$ from this value of 10.7 upward, to find the true largest value for which $K$ and $D$ are positive definite. By doing this it is obtained that for $\sigma = 90.1$ both $K$ and $D$ are positive definite, but for $\sigma = 83.1$, $D$ is not. This means that $\sigma = 90.1$ is used as the iteratively found upper bound. From this a $\sigma$ is chosen at around half of this upper bound, so $\sigma = 45$. The reasoning behind this choice will be further explained in Chapter 5.
In Figure 3.5 the values of the smallest eigenvalue and the eigenvalues with smallest magnitude (smallest absolute value of eigenvalues) can be seen for both matrices. The red lines give the smallest eigenvalues and eigenvalues of smallest magnitude for $K$, which correspond to the axis on the right. The blue lines for $D$, which correspond to the axis on the left. As can be seen in the figure is that as long as the eigenvalues are positive, the two lines of the same matrix coincide. For matrix $K$ this is the case for all sigma used here. The dashed green line denotes the upper bound obtained from calculation. The orange one gives the largest value for which both matrices are diagonally dominant. The purple dashed line gives the actual upper bound for diagonal dominance, as found above. In Figure 3.5a this is done for a large range, starting before the upper bound found from the calculation and ending after the bound obtained from iteration on positive definiteness. In Figure 3.5b the same is done, but on a smaller interval. There it is clearly visible that the smallest eigenvalue of $D$ becomes negative, around $\sigma = 90.1$, which was also the bound found from iteration. It is also visible that this is slightly larger the largest value which gives diagonal dominance.
From this the eventual value of $\sigma$ is chosen to be $\sigma = 45$. For this value the matrices $K$ and $D$ are visualised in Figure 3.6 and 3.7. There we see that the off-diagonal entries are very small compared to the diagonal elements. From a calculation the same can be obtained. On this field a minimum distance of 227.7 was found. This minimum distance gives the largest off-diagonal values when put in the equations which gave the terms for the matrices $K$ and $D$, Equations 2.8 and 2.27. We will let $[K]_m$ and $[D]_m$ be the terms that correspond to this. Now this gives

$$\begin{aligned} [K]_m &= \frac{1}{2\pi\sigma^2}\exp\left(-\frac{m^2}{2\sigma^2}\right) \\ &= \frac{1}{2\pi\cdot 45^2}\exp\left(-\frac{227.7^2}{2\cdot 45^2}\right) \\ &= 2.2\cdot 10^{-10} \\ [D]_m &= \frac{1}{8\pi\sigma^8}\exp\left(-\frac{m^2}{4\sigma^2}\right)\left(\frac{1}{2}\sigma^2 m^2 - m + 2\sigma^4\right) \\ &= \frac{1}{8\pi\cdot 45^8}\exp\left(-\frac{227.7^2}{4\cdot 45^2}\right)\left(\frac{1}{2}\cdot 45^2\cdot 227.7^2 - 227.7 + 2\cdot 45^4\right) \\ &= 2.4\cdot 10^{-10} \end{aligned} \tag{3.8}$$

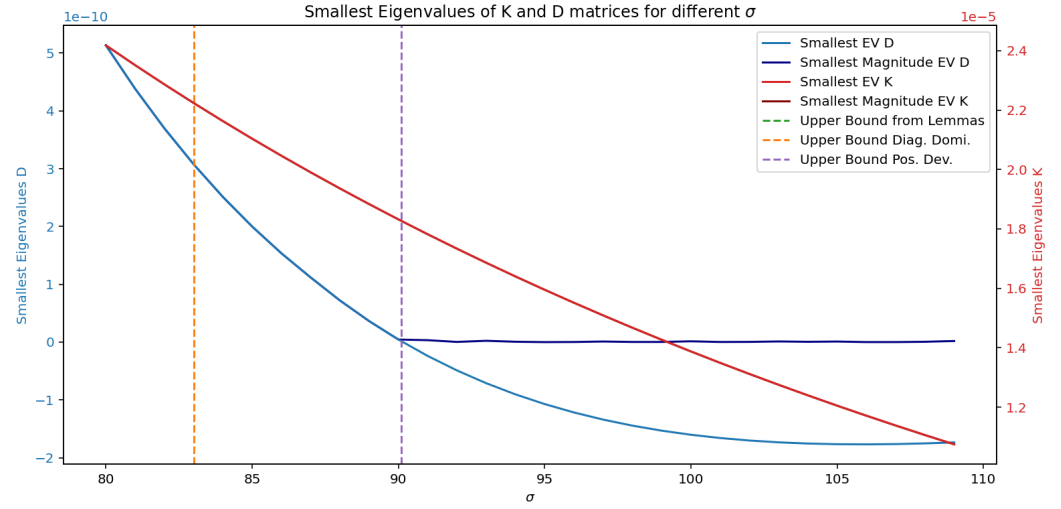
The same can be done for a diagonal term, which means that the distance is 0. This gives

$$\begin{aligned}
[K]_{p,p} &= \frac{1}{2\pi \cdot 45^2} \exp\left(-\frac{0^2}{2 \cdot 45^2}\right) \\
&= 7.85 \cdot 10^{-5} \\
[D]_{p,p} &= \frac{1}{8\pi \cdot 0^8} \exp\left(-\frac{227.7^2}{4 \cdot 0^2}\right)\left(\frac{1}{2} \cdot 45^2 \cdot 227.7^2 - 227.7 + 2 \cdot 0^4\right) \\
&= 1.94 \cdot 10^{-8}
\end{aligned} \tag{3.9}$$

This further confirms the large difference between the diagonal and of diagonal entries.

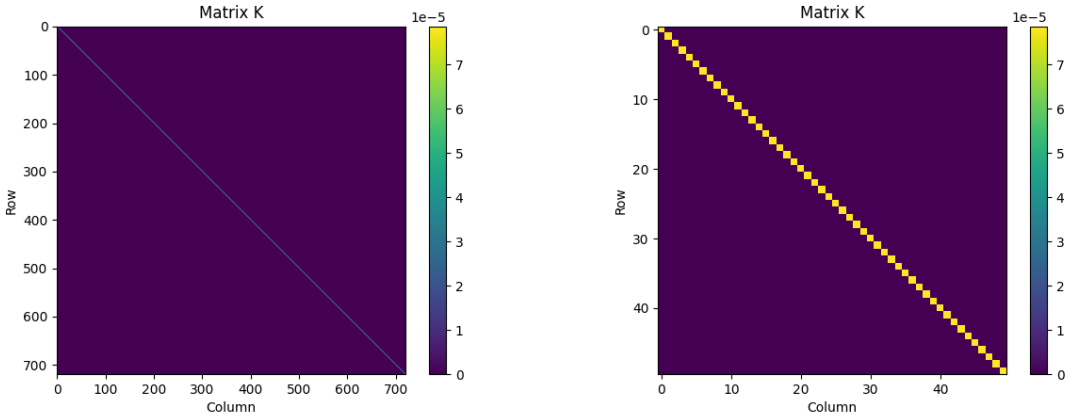



(a)




(b)

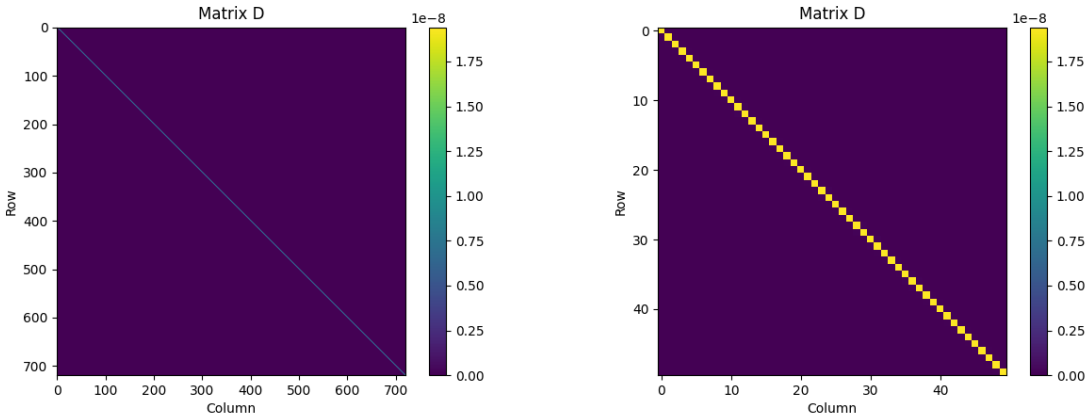
Figure 3.5: Smallest Eigenvalues for different $\sigma$ on a (a) large scale and (b) further zoomed in




(a) Full matrix

(b) Zoomed in on first 50 rows and columns

Figure 3.6: Visualisation of $D$-matrix for Montfrin 2021 with $\sigma = 45$




(a) Full matrix

(b) Zoomed in on first 50 rows and columns

Figure 3.7: Visualisation of $D$-matrix for Montfrin 2021 with $\sigma = 45$

# Chapter 4

# Finding an optimal value for the $\eta$ parameter

## 4.1 Finding $\eta$ using cross-validation

In Chapter 2 a method to find an approximation of the field effect, here often denoted $\gamma$, was discussed. This method had a parameter about which nothing was know yet, denoted by $\eta$. For this parameter an optimal value is tried to be found.
In the previous chapter a good way to choose $\sigma$ was proposed. This is one of two parameters needed to accomplish the goal that is wanted to be accomplished. In Chapter 1 the Euler-Lagrange equation was introduced, in Equation 1.17, on which was expanded upon in Chapter 2 using the kernel method. This gave

$$(K^T R^T RK + \eta D)\gamma = -K^T R^T R\mathbf{E}, \tag{4.1}$$

which was already found in Equation 2.5. Now using the $\sigma$ that is chosen it is possible to form the matrices $K$ and $D$. This leaves one parameter, the $\eta$, which must be found. The optimal value for this parameter will be found using cross-validation. For this the matrix $R$ will be split in a training and a validation part. This split can be made in three different ways. In each batch three plots were being compared to a fourth plot (in the way that it was used before plot 2, 3 and 4 were compared to plot 1). One of these three is taken and used as validation, the other two will be used as training. An example in which matrix $R$ can be split can be seen here, where plot 3 is used as validation. There the second row is taken out and forms the validation matrix, resulting in a $2 \times 4$ training matrix and a $1 \times 4$ validation matrix.

$$R_{bt} = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 1 & 0 & 0 & -1 \end{bmatrix}, \quad R_{bv} = \begin{bmatrix} 1 & 0 & -1 & 0 \end{bmatrix}. \tag{4.2}$$

Of course this does not necessarily have to be the second row; it can also be the other two. This means that there are three ways to split the matrices in training and validation. All three will be used here, giving three processes of cross-validation.
For both validation and training Equation 4.1 can be solved for some chosen value of $\eta$. Using this an optimal value for $\eta$ can be found. To find this optimal value for $\eta$ there is being iterated over different values of $\eta$ and each time Equation 4.1 is solved for the training part. This gives a solution $\gamma_\eta$, for which it is desired to know whether this is a good solution. To find out whether a solution is a good solution, and to eventually find the optimal solution, residuals are used.
For this method two residuals are calculated, one for training denoted by $\rho_t = \rho(\mathbf{E}, \gamma_\eta, R_{bt})$ and one for validation $\rho_v = \rho(\mathbf{E}, \gamma_\eta, R_{bv})$, so for training $R_{bt}$ is used and for validation $R_{bv}$. These residuals are obtained from the following formula

$$\rho(\mathbf{E}, \gamma_\eta, R) = \frac{|R\mathbf{E} + RK\gamma_\eta|}{|R\mathbf{E}|}, \tag{4.3}$$

where $|\cdot|$ gives the $\ell^2$-norm.
The residuals for both training and validation are calculated using the $\gamma_\eta$ found from solving the system

$$(K^T R_{bt}^T R_{bt} K + \eta D)\gamma_\eta = -K^T R_{bt}^T R_{bt}\mathbf{E}, \tag{4.4}$$

which is exactly the same as the one in Equation 4.1, but with $R_{bt}$ instead of $R$. This is done three times, once for each cross-validation process obtained from the different ways in which $R$ can be split. This means that for each value of $\eta$ three different $\gamma_\eta$ are found, resulting in three different values for both the training and validation residual. From these three values the averages are calculated, resulting in a average residual for both training and validation. From doing this it will become clear for which value of $\eta$ the average residual of the validation is the smallest. The $\gamma_\eta$ that corresponds to this $\eta$ then should give the best estimate of the field effect.

In Figure 4.1 the residuals for both training and validation can be seen for the field in Montfrin in 2021. The blue lines show the training, the orange ones the validation. Both training and validation are calculated three times, since each of the rows of $R_b$ is used as validation once. Each one of these times is shown by the dashes lines. The two solid lines give the average of the three residuals, one for each of the averages of validation and training. As can be seen the training residual starts of very small, when $\eta$ is very small as well. This is exactly what is to be expected when looking at Equation 4.1. If there $\eta$ is close to 0, $\gamma$ becomes such that it perfectly estimates the values of $\mathbf{E}$ when using $R_{bt}$. This is because if $\eta$ is small the influence of $\eta D$ is very small. The problem there is that it does not necessarily estimate $\mathbf{E}$ as good when using $R_{bt}$ instead. The optimal value for $\eta$ is found to be 0.05, which gives a residual of 0.776. Both are indicated by the green dashed lines.
For the obtained optimal $\eta$ the total residual can be calculated by using the full matrix $R$ in Equation 4.3, instead of the split training or validation matrix. This is the residual of the estimate of the entire vector of data $\mathbf{E}$ that is obtained from $\gamma$. This total residual gives how much of the difference in yield can be attributed to spatial effect and how much to statistical error. For the field in Montfrin in 2021 this total residual is 0.18 for this optimal value of $\eta$ on this field. This means that 82 % of the difference in yield can be attributed to spatial effect. The other 18 % can not be explained by this model and is thus attributed to statistical error.
The values of optimal $\eta$, minimal validation residual and the corresponding total residual have been found for the other fields as well and are shown in Table 4.1.
Some puzzling results can be observed in the rows of the table corresponding to SPNA 2020 and Veenklooster 2020. There the values of $\eta$ are incredibly small which means that the obtained $\gamma$ is a very good approximation of the training data. At the same time, the residual graph obtained for these fields is not smooth around the optimal $\eta$, as can be seen in Figure 4.3, where the residuals for SPNA in 2020 can be seen. The results for Veenklooster 2020 are similar. The validation residual for these fields hardly changes with $\eta$ and oscillates on the level of numerical noise. This behaviour might have to do with the canopy data or the geometry of the fields. Whatever the reason, these fields represent obvious outliers and will not be considered further in this project. Figures for the residuals of the other seven fields are shown in Appendix B.

## 4.2 Further justification of the choice of $\sigma$

In Section 3.2.2 it was mentioned that $\sigma$ must not be chosen too close to the upper bound. This has to do with the residuals that are obtained from cross-validation. If $\sigma$ were to be chosen closer to its upper bound the residuals get much worse without improving the interpretability. For some fields it even turns out to be impossible to find an optimal value for $\eta$. The validation residual will stay at or be close to 1, which means that the total residual becomes large as well. The interpretation obtained from this would be that all, or nearly all, of the difference in yield data has to be attributed to statistical error, and that the obtained field effect does not say anything. An example of this can be seen in Figure 4.2. There the residuals of the field in Montfrin in 2019 is visible with a $\sigma$ chosen close to the upper bound obtained from Table 3.1. This upper bound was 19.4, so $\sigma$ is chosen to be 18. In this figure it can be seen that while the validation residual stays larger than 1, the training residual grows to 1 as well. This means that before an optimal value for $\eta$ is found, the total residual is already is way to large. For example from this Figure 4.2 a total residual of 0.98 is obtained, which means the obtained field effect only explains 2% of the observed difference, with the rest coming from statistical errors. A field effect which only explains 2% of the differences is of course not very useful.

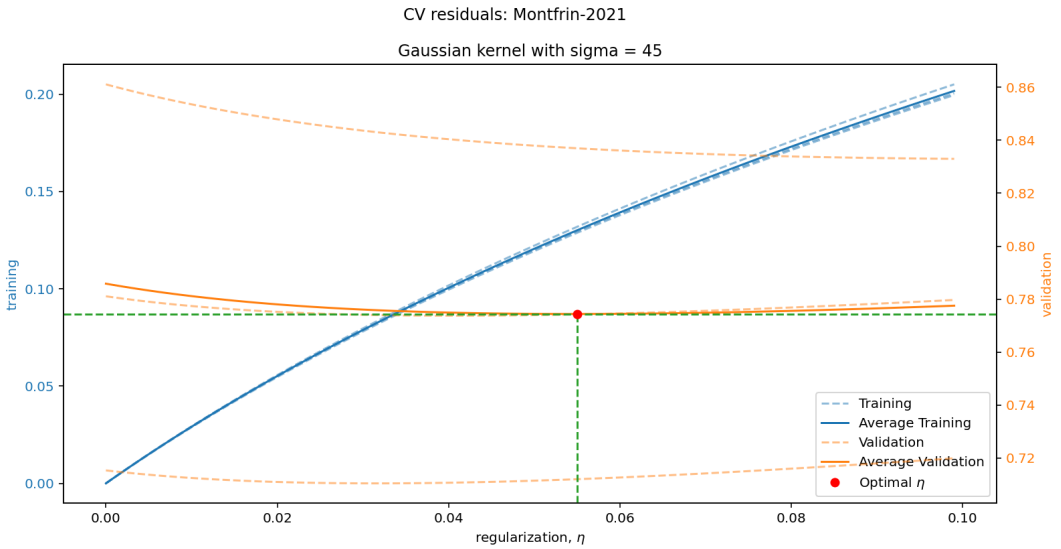



Figure 4.1: Residuals of training and validation of the field in Montfrin in 2021

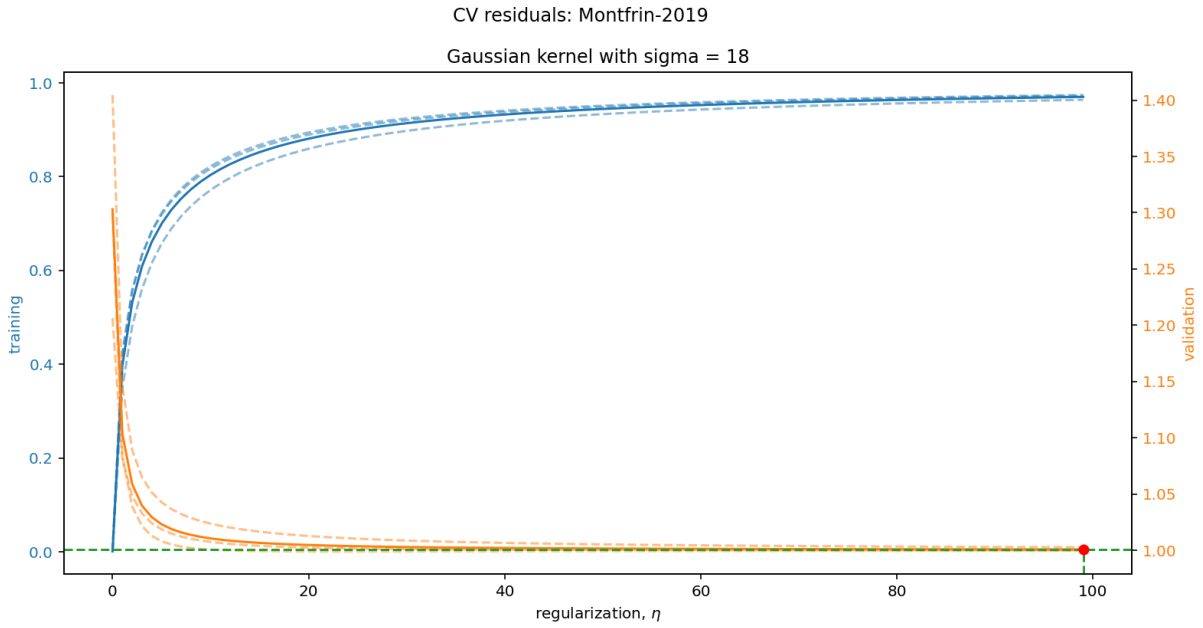



Figure 4.2: Residuals of training and validation of the field in Montfrin in 2019 with a to large $\sigma$

| Field | Year | Optimal $\eta$ | Validation Residual | Total Residual |
|---|---|---|---|---|
| Montfrin | 2019 | 3.4 | 0.96 | 0.70 |
| | 2020 | 4.2 | 0.93 | 0.58 |
| | 2021 | 0.052 | 0.77 | 0.18 |
| SPNA | 2019 | 0.068 | 0.80 | 0.10 |
| | 2020 | $10^{-15}$ | 0.77 | $2.6 \cdot 10^{-15}$ |
| | 2021 | 0.44 | 0.89 | 0.31 |
| Veenklooster | 2019 | 3.1 | 0.87 | 0.22 |
| | 2020 | $1.45 \cdot 10^{-14}$ | 0.70 | $2.2 \cdot 10^{-14}$ |
| | 2021 | 0.30 | 0.91 | 0.49 |

Table 4.1: Optimal values of $\eta$ and their residuals

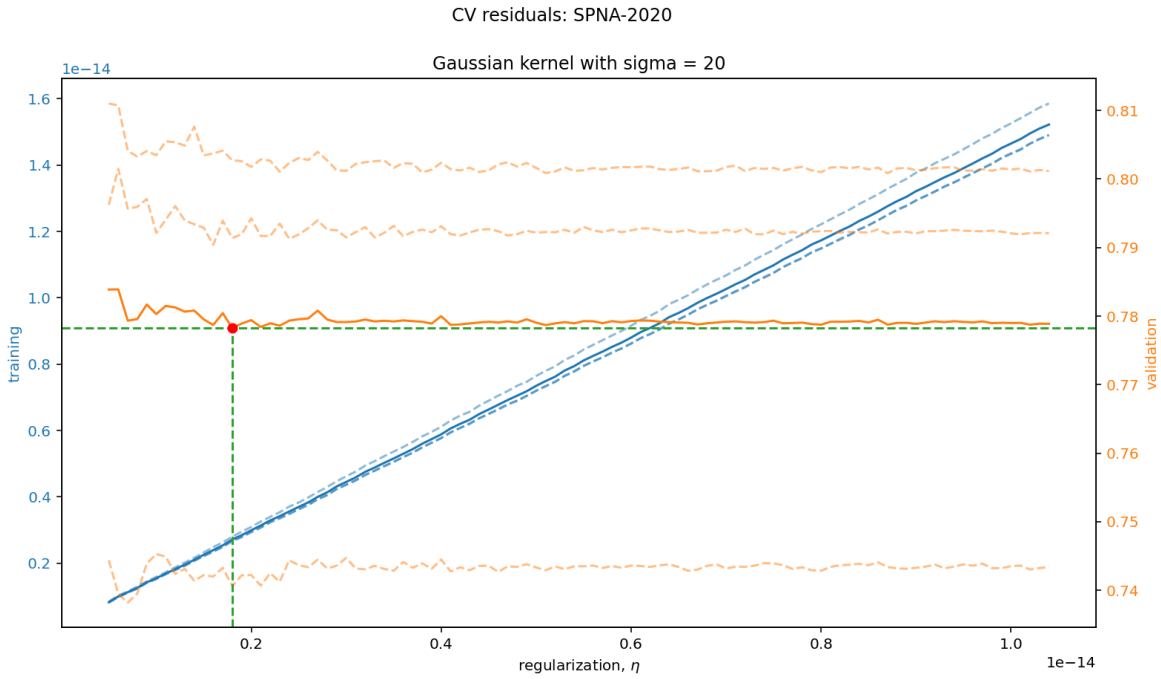



Figure 4.3: Residuals of training and validation of the field in SPNA in 2020

# Chapter 5

# Results

## 5.1 Field effect

With values for $\eta$ and $\sigma$ found, everything is obtained to solve Equation 4.1 for $\gamma$. The $\gamma$ gives the field effect in the positions in which the data was known, so the $\gamma_p$ which we saw before in Chapter 1. Using these the field effect in any position on the field can be obtained from

$$\gamma(\mathbf{r}) = \sum_{p=1}^{P} \gamma_p V_p(\mathbf{r}). \tag{5.1}$$

As we saw before in Chapter 2 these $V_p$ can be obtained easily from the kernel with

$$V_p(\mathbf{r}) = \kappa(||\mathbf{r} - \mathbf{r}_p||). \tag{5.2}$$

This will be used to find the field effect and virtually replant the data on the different fields. Here two fields will be discussed more extensively, one with a small total residual in Table 4.1 and one with a larger total residual in Table 4.1. For the one with the a small total residual much of the differences in canopy can be attributed to field effect and thus the field effect obtained on this field should give a lot of information. For a field with a large total residual this is not the case. There most of the differences in canopy of plots of the same batch must be attributed to statistical errors, which means that the information obtained from the field effect gives less information that is useful.
For the fields that are not further discussed here the results can be found in Appendix B. The field in Montfrin in 2021 is chosen as the field with small total residual. From the fields where nothing strange happens to the values of $\eta$, so all fields except for SPNA 2020 and Veenklooster 2020, Montfrin 2021 has the second smallest total residual. The only nicely working field which has a smaller total residual is SPNA 2019. Montfrin 2021 is chosen over SPNA 2019 because Montfrin 2021 has already been used in all other examples and because the visualisation of Montfrin looks better than the one of SPNA.
The field that is chosen as the field with large total residual is the field in Montfrin in 2019. This is the field with the largest total residual, as obtained in Table 4.1.

### 5.1.1 Replanting with small total residual

Using Equation 5.1 the field effect can be found easily in any position on the field. To visualise this function a uniform grid is made on the field and in each point of this grid $\gamma(\mathbf{r})$ is calculated. For the field in Montfrin in 2021, where the total residual has a relatively small value of 0.18, the field-effect function can be seen in Figure 5.1. There a blue color means that the plot is effected positively by field effect and requires a negative correction. Thus the data observed, which is visible in Figure 3.4, for such a plot is higher than the expectation for that batch. This means that such a plot after spatial correction will have a smaller canopy. A red color means the exact opposite: the field effect influences these plots negatively and thus after correction these plots will have a larger canopy.
Now that the field effect is found the virtual replanting can be done. Using this virtual replanting the mean canopy for each plot is transformed to a value that would have been observed if this plot was located at a

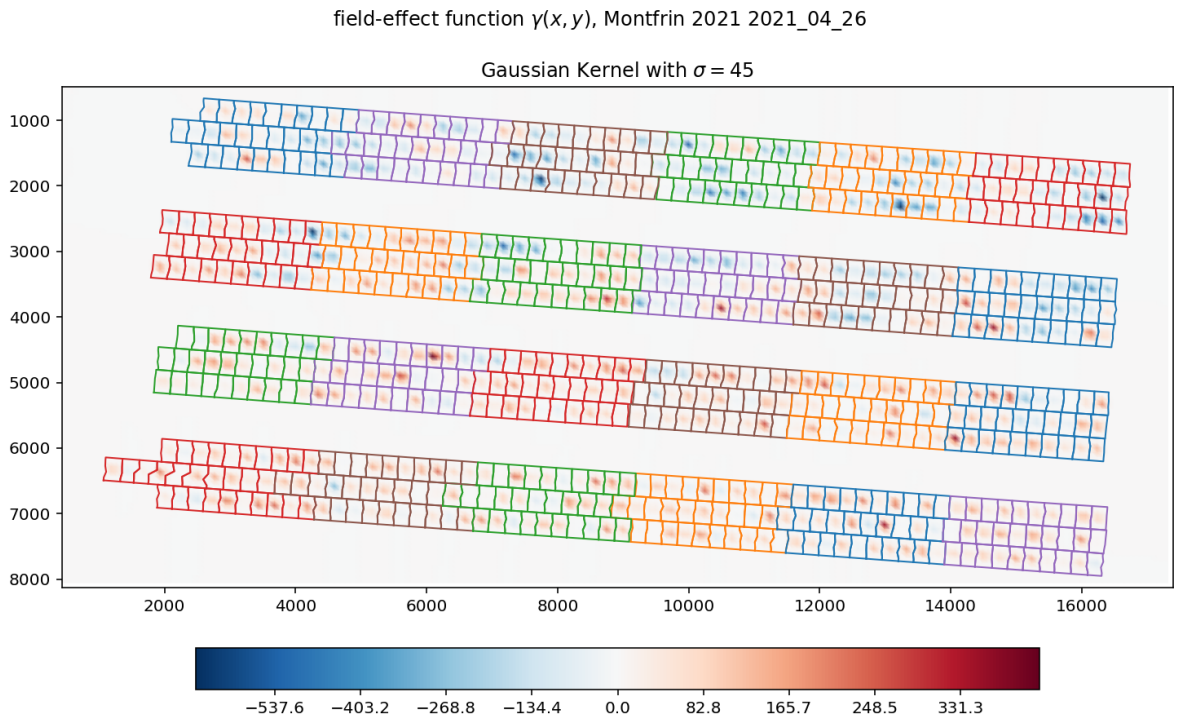
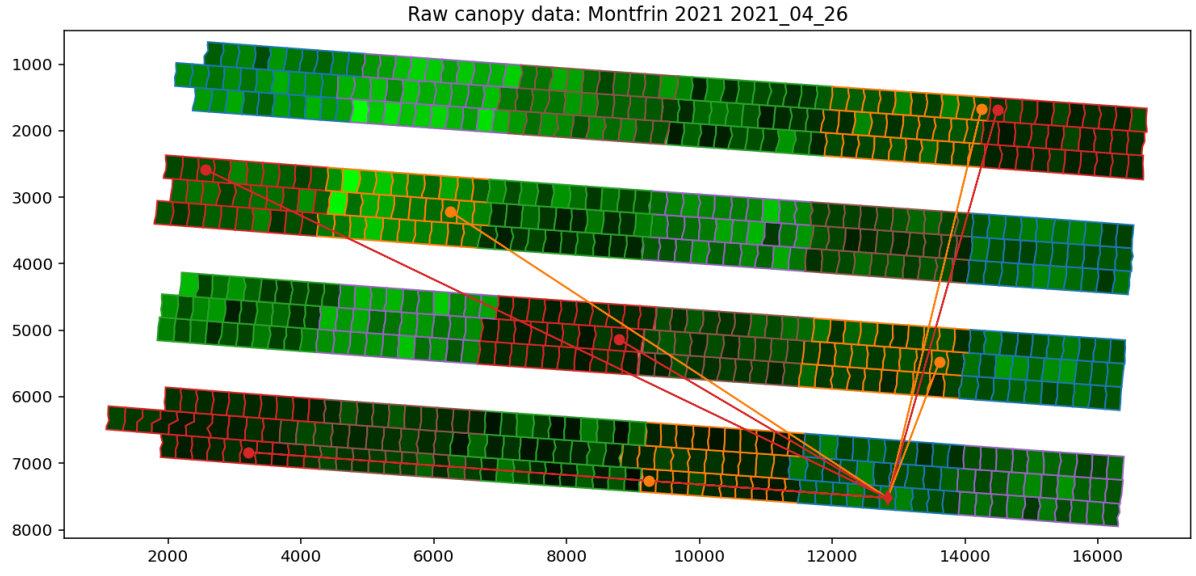
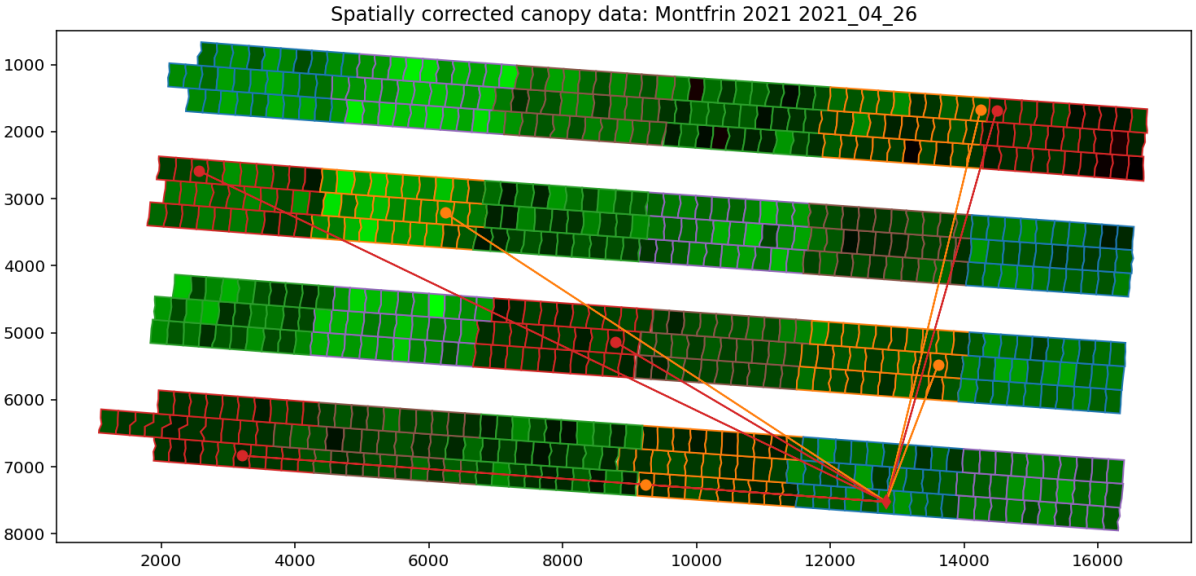



Figure 5.1: Field Effect Function on the field in Montfrin in 2021

different location. One plot is chosen as the target location where all plants are virtually moved to. This gives an estimate for the field as if all plants were planted in one location. In this one location the field effect is a fixed value. Thus if all plants were planted in this one location and there would be no statistical noise, then there would be no field effect observed. This means that in the virtually replanted data the differences between the plots of the same batch are mostly caused by statistical errors.
The replanting of a plot is done by taking the difference between the field effect of said plot and the target location. This gives the difference in yield that can be attributed to spatial effect. By subtracting this found difference from the obtained data of the plot that is replanted corrected data can be found. This corrected data gives the canopy which would be found if the plot that is being replanted were planted at the target location.
Everything is now found to do the replanting that was desired to do. In Figure 5.2a the original data for Montfrin 2021 can be seen once again, just like in Figure 3.4. Now four plots are marked with an orange dot and four with a red dot. The plots with a dot of the same color are planted with plants from the same batch, batch 31 for the orange dots and batch number 65 for the red dots. For these batches the original and corrected data will be observed. The plot marked by the red diamond is the target location, to which all plants are being replanted. In Figure 5.2b the data can be seen after spacial correction. The same plots are marked with orange as in Figure 5.2a. Between the figures no immediate difference is seen, but we can compare some of the obtained values for the marked batches.
In Table 5.1 on the right the values for batch 65 can be seen before and after replanting. For this batches the original values and the values after replanting behave exactly in the way one would expect. The small values have increased and the large values have decreased, dragging all values towards a average values.
For the other batch, batch number 31, this is not the case. There the large values are corrected to even larger values. This is unexpected behaviour and might have to do with the model not functioning perfectly or with statistical errors.




(a)




(b)

Figure 5.2: Original (a) and spatial corrected (b) data on the field in Montfrin in 2021 with plots marked for replanting and marked target location

| Index | Original Canopy | Spatial Corrected Canopy |
|---|---|---|
| 1 | 883.4 | 762.0 |
| 2 | 363.7 | 386.2 |
| 3 | 1506.5 | 1622.3 |
| 4 | 728.4 | 741.6 |

| Index | Original Canopy | Spatial Corrected Canopy |
|---|---|---|
| 1 | 322.4 | 382.0 |
| 2 | 456.2 | 406.0 |
| 3 | 734.2 | 693.6 |
| 4 | 303.2 | 341.1 |

Table 5.1: Original and spatially connected data for batch 31(left) and batch 65 (right)

### 5.1.2 Replanting with a large total residual

It is also interesting to see what happens with the model when applied to a field with a performance that is not as good. Therefore the same things applied to Montfrin 2021 will be applied to Montfrin 2019.

In Figure 5.3 the field effect on this field can be seen. This figure shows the same as Figure 5.1, with a blue color meaning a positive influence from field effect and a red color meaning a negative influence from field effect. Here it can immediately be seen that the values of the influences of the field are small compared to the ones found in Figure 5.1. Where in Figure 5.1 this ranged from -550 to 350, in Figure 5.3 it only ranges from -150 to 100. This might have to do with the fact that the field effect explains a far smaller portion of the differences in data and a larger portion is explained by statistical error
In Figure 5.4 the original data and the replanted data can be seen. Again two batches are marked of which the values can be seen in Table 5.2. Now the batches which are used are batch 60, in orange in the figures and in the table on the left, and batch 65, in red in the figures and in the table on the right.
For both batches we see that for most of the indices the corrected data is greater than the original data. In seven of the eight plots given here this is the case and this is also observed in other batches. This corresponds with Figure 5.3, where more red is visible than blue, meaning that the corrections mostly are positive.

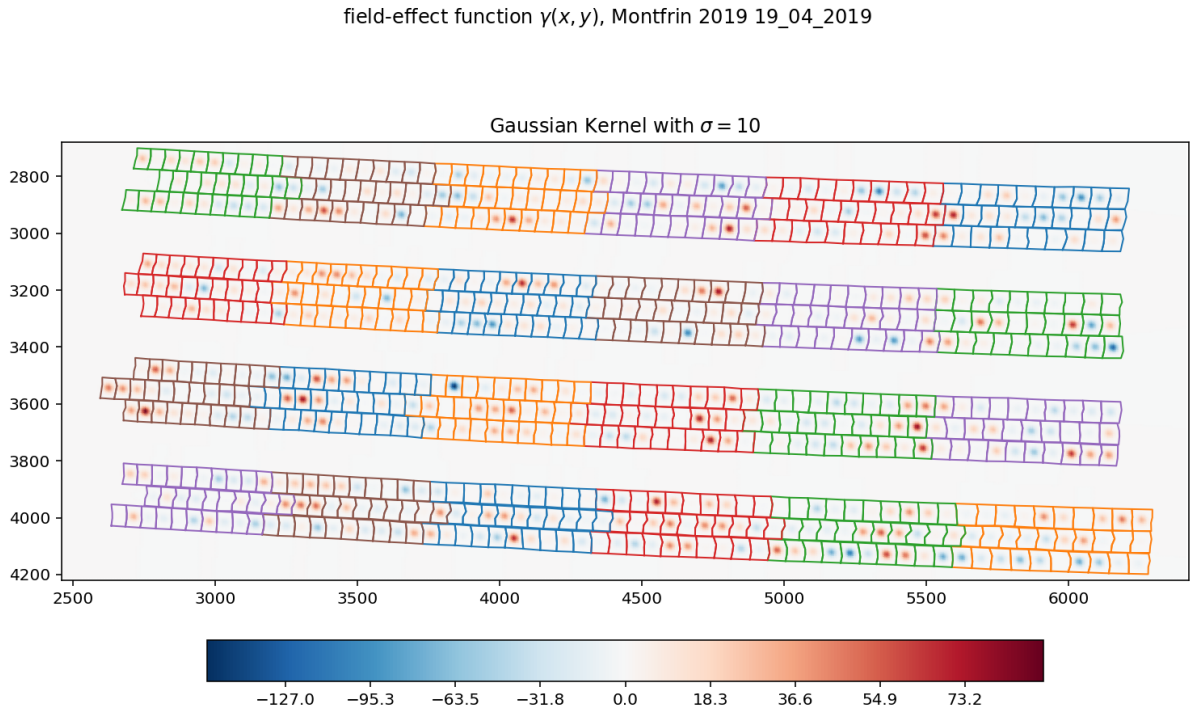
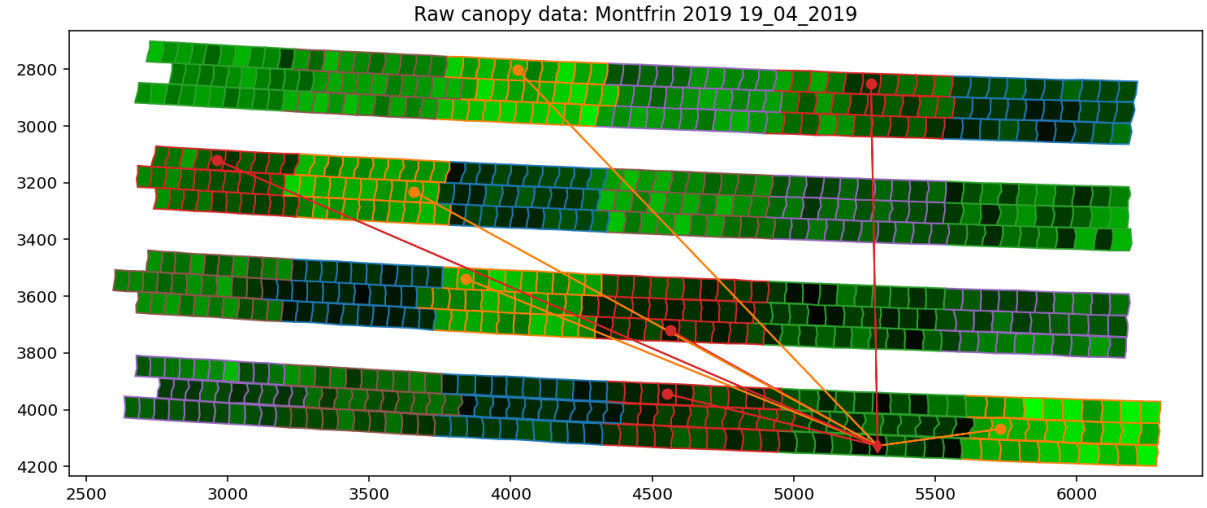



Figure 5.3: Field Effect Function on the field in Montfrin in 2019

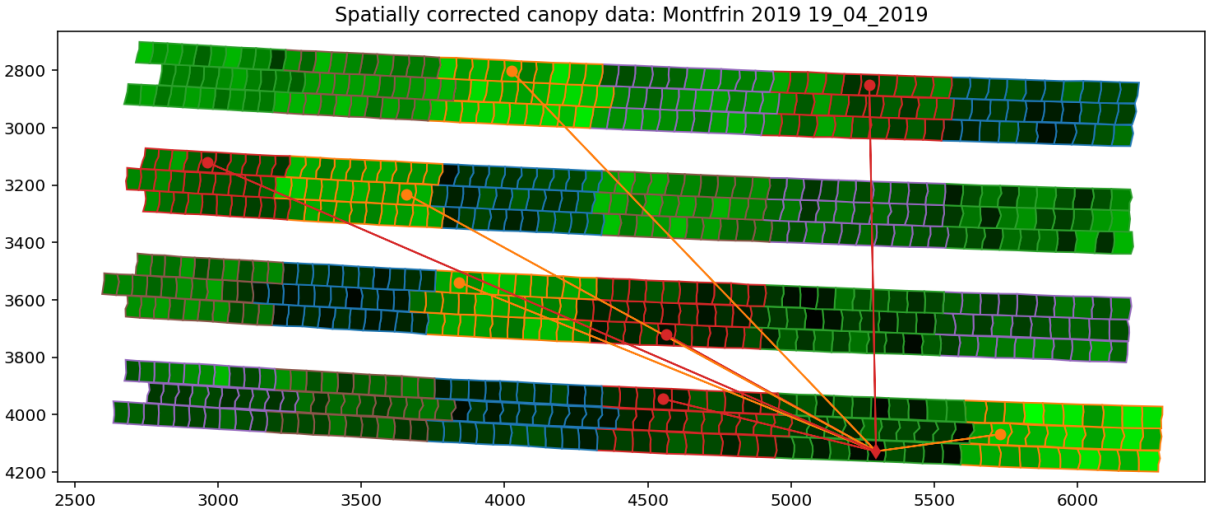
| Index | Original Canopy | Spatial Corrected Canopy |
|---|---|---|
| 1 | 1097.2 | 1187.7 |
| 2 | 1025.2 | 971.4 |
| 3 | 896.7 | 1000.5 |
| 4 | 986.9 | 1080.9 |

| Index | Original Canopy | Spatial Corrected Canopy |
|---|---|---|
| 1 | 168.7 | 346.4 |
| 2 | 105.4 | 209.1 |
| 3 | 338.0 | 450.7 |
| 4 | 247.4 | 289.0 |

Table 5.2: Original and spatially connected data for batch 60(left) and batch 65 (right)




(a)




(b)

Figure 5.4: Original (a) and spatial corrected (b) data on the field in Montfrin in 2019 with plots marked for replanting and marked target location

## 5.2 Correlation

In the previous section we saw some values for original and corrected data already, but it is desired to know more about these values to be able to get a conclusion from it.
Something else that can be used to find out how well the model performs is researching the correlation between the data on the different fields. On fields that are planted in the same year the same batches are planted, so there the correlation is expected to give some good information. This correlation is calculated both between the original data of the fields and between the corrected data obtained using the kernel method. The way to calculate correlation is by using Pearson correlation, where the correlation between two vectors of data, $x$ and $y$, can be found using

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - \left(\sum y_i\right)^2}}. \tag{5.3}$$

In Table 5.3 the correlations for the different years can be seen for the original data set. In Table 5.4 the correlations for the corrected data can be seen. Obviously in all tables the values on the diagonal are equal to 1, since the correlation of a vector with itself is one. The tables also are symmetric, as for two vectors the correlations $r(x,y) = r(y,x)$.
From these tables a couple of observations can be made that are unexpected. The first of this is that there is negative correlation between the Montfrin 2021 and the other two fields in 2021. This is unexpected, as a batch that performs well on one field is expected to perform well on another field as well. This negative correlation would suggest that the better a batch performs on Montfrin in 2021, the worse it does on the other fields.
Something else that is unexpected is that the correlation does not improve for the corrected data. The correlation in most cases even decreases after the correction. This is not the behaviour that is expected, as the correction for field effect should give data that is closer to the expectation of the batches. This might suggest that the model that is used does not work perfectly.

| **2019** | Montfrin | SPNA | Veenklooster |
|---|---|---|---|
| Montfrin | 1 | 0.528 | 0.579 |
| SPNA | 0.528 | 1 | 0.543 |
| Veenklooster | 0.579 | 0.543 | 1 |
| **2020** | Montfrin | SPNA | Veenklooster |
| Montfrin | 1 | 0.602 | 0.273 |
| SPNA | 0.602 | 1 | 0.176 |
| Veenklooster | 0.273 | 0.176 | 1 |
| **2021** | Montfrin | SPNA | Veenklooster |
| Montfrin | 1 | -0.093 | -0.174 |
| SPNA | -0.093 | 1 | 0.658 |
| Veenklooster | -0.174 | 0.658 | 1 |

Table 5.3: Correlations of original data in three different years, 2019 (top) 2020 (middle) and 2021 (bottom)

| **2019** | Montfrin | SPNA | Veenklooster |
|---|---|---|---|
| Montfrin | 1 | 0.485 | 0.531 |
| SPNA | 0.485 | 1 | 0.543 |
| Veenklooster | 0.531 | 0.496 | 1 |
| **2020** | Montfrin | SPNA | Veenklooster |
| Montfrin | 1 | 0.598 | 0.185 |
| SPNA | 0.598 | 1 | 0.138 |
| Veenklooster | 0.185 | 0.138 | 1 |
| **2021** | Montfrin | SPNA | Veenklooster |
| .0Montfrin | 1 | -0.091 | -0.215 |
| SPNA | -0.091 | 1 | 0.642 |
| Veenklooster | -0.215 | 0.642 | 1 |

Table 5.4: Correlations of corrected data in three different years, 2019 (top) 2020 (middle) and 2021 (bottom)

# Conclusions and recommendations

In this project ways to find field effects on fields of crops using yield or canopy data were researched. For this a model was introduced, the replanting model, which compares the values of a phenotype parameter of plots with the plants of the same production origin at different locations in the field. The main goal of this project was to apply this model in a way such that a function for the field effect could be found that does not depend on a mesh of the field. The kernel method with the Gaussian kernel function allowed to find the field effect function without the use of a mesh of the field. An explicit expression for the elements of the stiffness matrix corresponding to the Gausssian kernel was derived. This matrix is employed as a gerularization term for controlling the smoothness of the field-effect function.
In the process of finding the field effect some limitations where found. The choice of the kernel-function parameter $\sigma$ appears to be limited by an upper bound that makes sure that all matrices stay positive definite. Keeping the representation and stiffness matrices positive definite allowed to employ the regularization theory without modifications. However, without these constraints there might be ways to obtain better results.
The results for the field effects that where found were still far from perfect. In the spatially corrected data no clear improvement was visible and the inter-field correlation got worse after spatially correcting the data. This suggests that some parts of the method to find the field effect function are not working as expected and should be improved.
For example, the ways in which the optimal parameters $\sigma$ and $\eta$ are determined. Here this was done by finding a value for $\sigma$ for which the computations certainly could be done, since the $K$ and $D$ matrices remained positive definite, and then finding an optimal value for $\eta$. This could be done differently, for example, by fixing $\eta$ and finding an optimal value of $\sigma$ or even by looking at $\sigma$ above the bound for positive definiteness.
Another recommendation is trying out other kernel functions. Another kernel function was already considered:

$$\kappa(\mathbf{r_p}, \mathbf{r_q}, \alpha) = \frac{1}{\left(1 + \frac{|\mathbf{r_p} - \mathbf{r_q}|^2}{\alpha}\right)^{\alpha}}. \tag{5.4}$$

The problem with this kernel turned out to be the construction of the $D$-matrix. The integral obtained for the entries of this matrix could not be evaluated explicitly. This kernel function could be promising if the entries of this matrix are found.
In any case, the kernel method presents a viable alternative for finding and visualising field effects without influence or even the need of a mesh.

# Appendix A

# Derivation of the Euler-Lagrange Equation

It is desired find an equation that finds a minimum for the functional $F_\eta(\gamma)$. This should find a $\gamma$ which satisfies

$$\frac{\partial F_\eta}{\partial \gamma_q} = 0, \tag{A.1}$$

for $q \in 1, \dots P$.
For this we first derive the partial differential of $F_\eta(\gamma)$.

$$\begin{aligned}
\frac{\partial F_\eta}{\partial \gamma_q} &= \frac{1}{2}\frac{\partial}{\partial \gamma_q}\left[(RK\gamma + R\mathbf{E})^T (RK\gamma + R\mathbf{E})\right] \\
&\quad + \eta \int_\Omega \frac{1}{2}\frac{\partial}{\partial \gamma_q}\left[\left(\sum_{p=1}^{P} \gamma_p \nabla v_p(\mathbf{r})\right)\cdot\left(\sum_{p=1}^{P} \gamma_p \nabla v_p(\mathbf{r})\right)\right] d\mathbf{r} \\
&= \left[\frac{\partial}{\partial \gamma_q}(RK\gamma + R\mathbf{E})^T\right](RK\gamma + R\mathbf{E}) \\
&\quad + \eta \int_\Omega \sum_{p=1}^{P} \gamma_p \nabla v_p(\mathbf{r}) \cdot \frac{\partial}{\partial \gamma_q}\left(\sum_{p=1}^{P} \gamma_p \nabla v_p(\mathbf{r})\right) d\mathbf{r} \\
&= [(RK)^T]_{q,:}(RK\gamma + R\mathbf{E}) + \eta \int_\Omega \sum_{p=1}^{P} \gamma_p \nabla v_p(\mathbf{r}) \cdot \left(\sum_{p=1}^{P} \delta_{p,q} \nabla v_p(\mathbf{r})\right) d\mathbf{r} \\
&= [(RK)^T]_{q,:}(RK\gamma + R\mathbf{E}) + \eta \int_\Omega \sum_{p=1}^{P} \gamma_p \nabla v_p(\mathbf{r}) \cdot \nabla v_q(\mathbf{r})\, d\mathbf{r} \\
&= [(RK)^T]_{q,:}(RK\gamma + R\mathbf{E}) + \eta \sum_{p=1}^{P} \gamma_p \int_\Omega \nabla v_p(\mathbf{r}) \cdot \nabla v_q(\mathbf{r})\, d\mathbf{r}
\end{aligned} \tag{A.2}$$

Now here the $[(RK)^T]_{q,:}$ denotes the $q$-th row of the matrix. By putting these all together the Euler-Lagrange equation of Chapter 1 is obtained:

$$(K^T R^T RK + \eta D)\gamma_b = -R^T R\mathbf{E}, \tag{A.3}$$

The $D$-matrix is the stiffness matrix with the elements

$$[D]_{p,q} = \int_\Omega \nabla v_p(\mathbf{r}) \cdot \nabla v_q(\mathbf{r})\, d\mathbf{r}. \tag{A.4}$$

# Appendix B

# Extra Figures

This appendix includes all figures for fields which were not already included in the report. This means for most fields the graph of the residuals, the field effect the original data and the corrected data. For some fields only a selection of those is included here, as the others were already included in the report. None of the figures for Montfrin 2021 are included at all, as these were all visible in the main report.

## B.1 Montfrin 2019

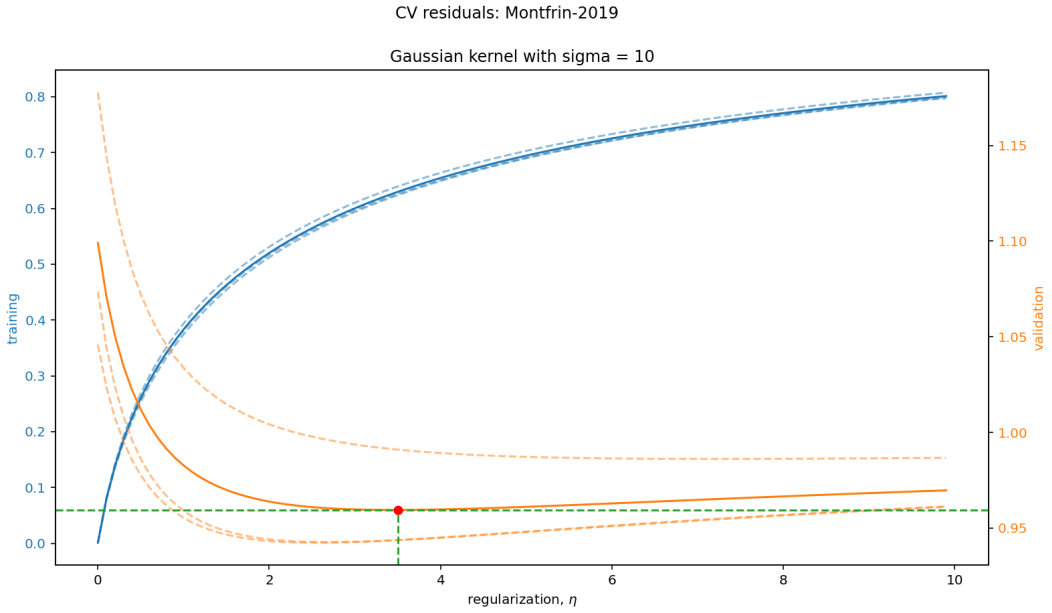



Figure B.1: Residuals of training and validation of the field in Montfrin in 2019

The field effect, the original data and the corrected data all were visible in the main report so they won't be included here.

## B.2 Montfrin 2020

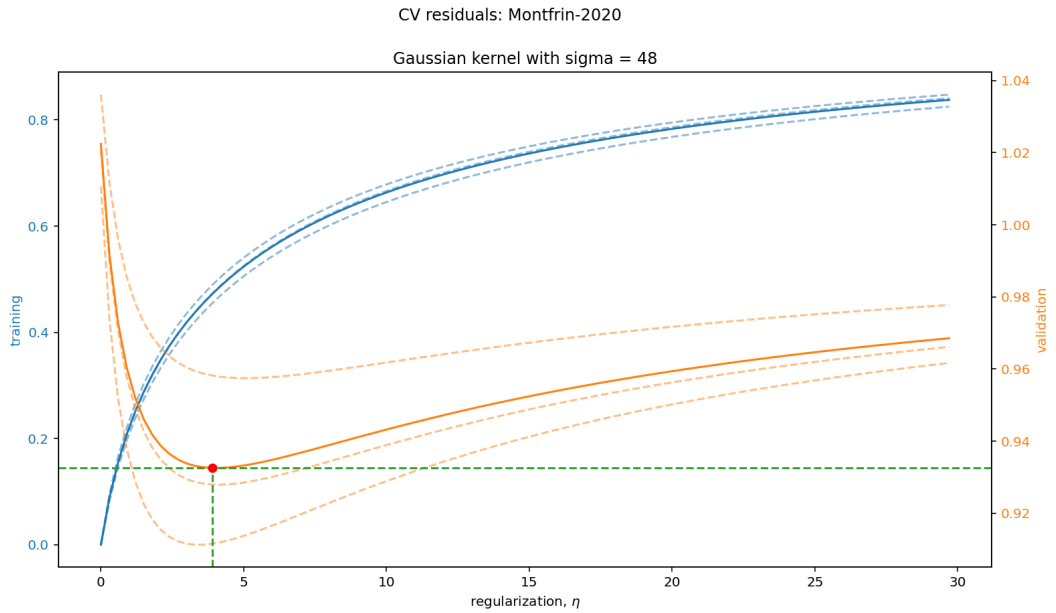



Figure B.2: Residuals of training and validation of the field in Montfrin in 2020

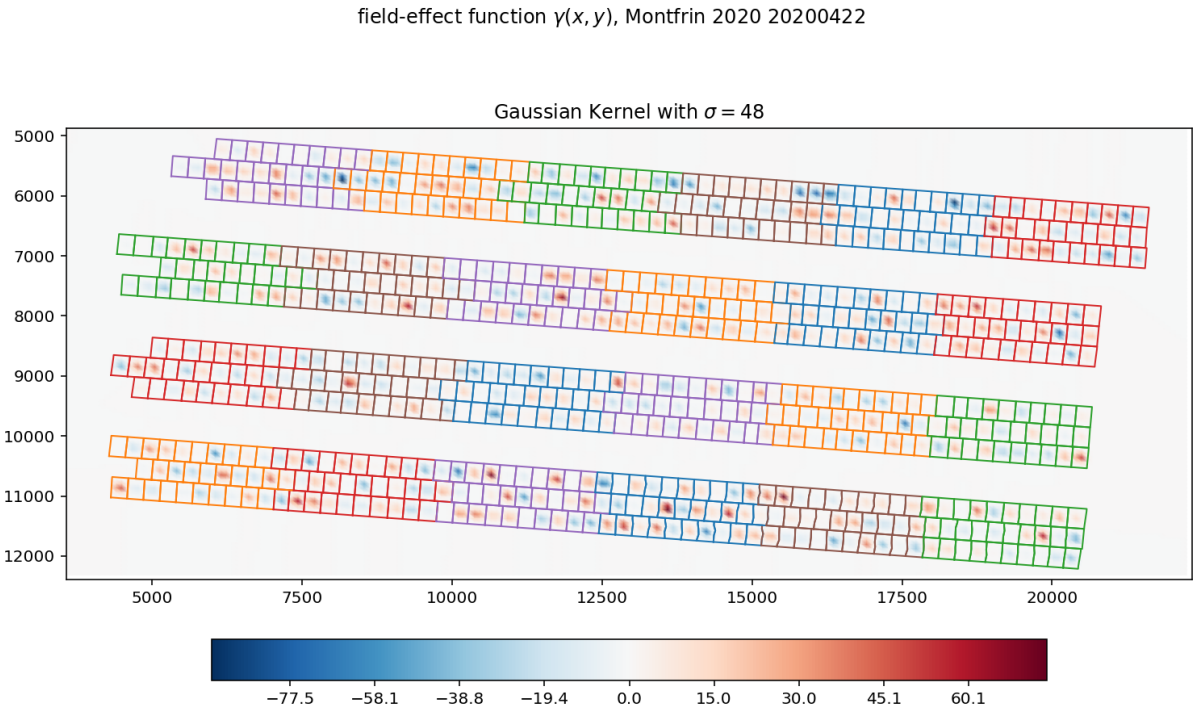
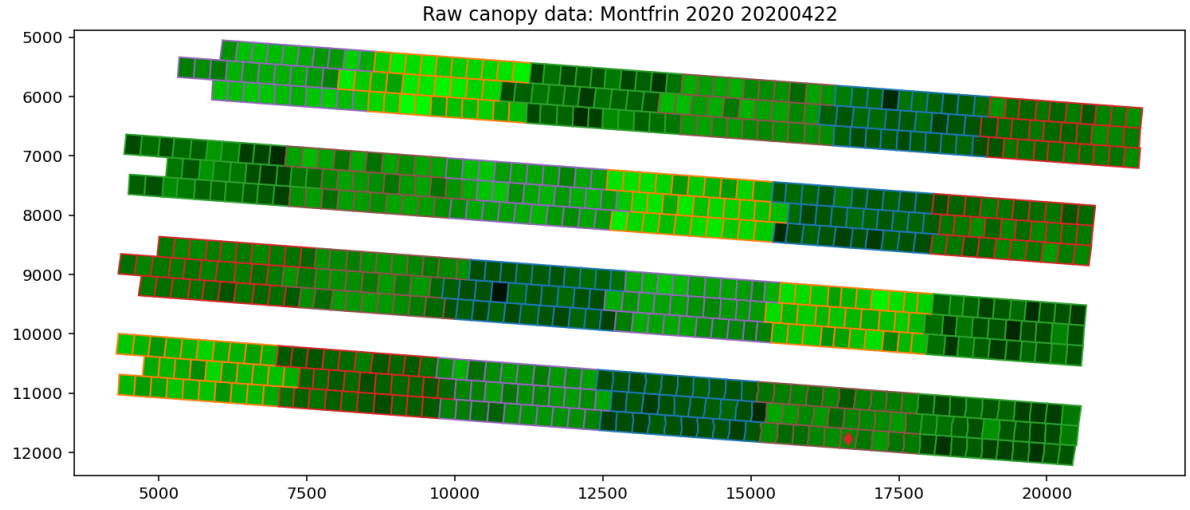
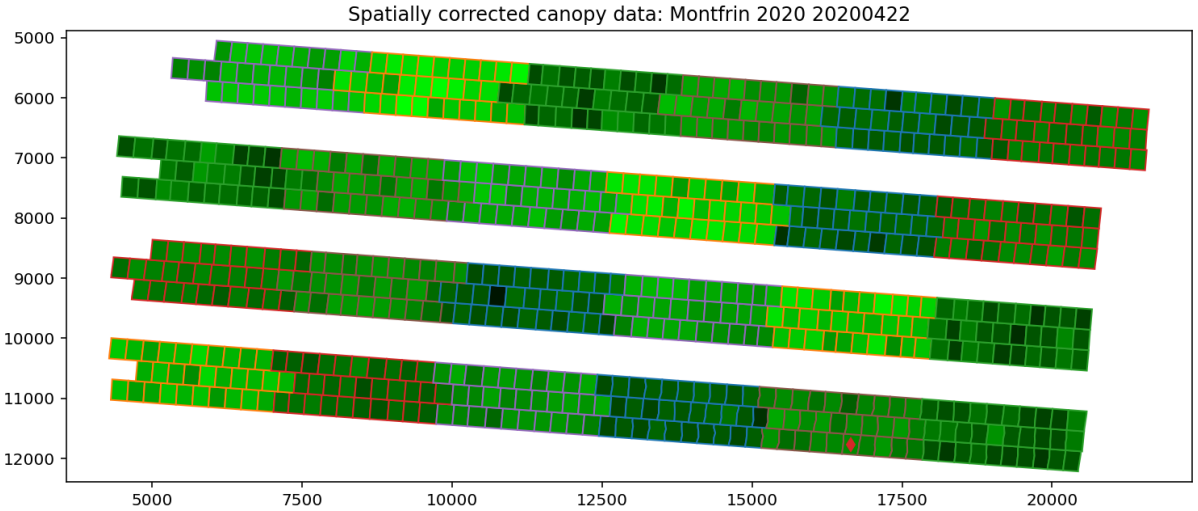



Figure B.3: Field Effect Function on the field in Montfrin in 2020




(a)




(b)

Figure B.4: Original (a) and spatial corrected (b) data on the field in Montfrin in 2020 with marked target location

## B.3 SPNA 2019

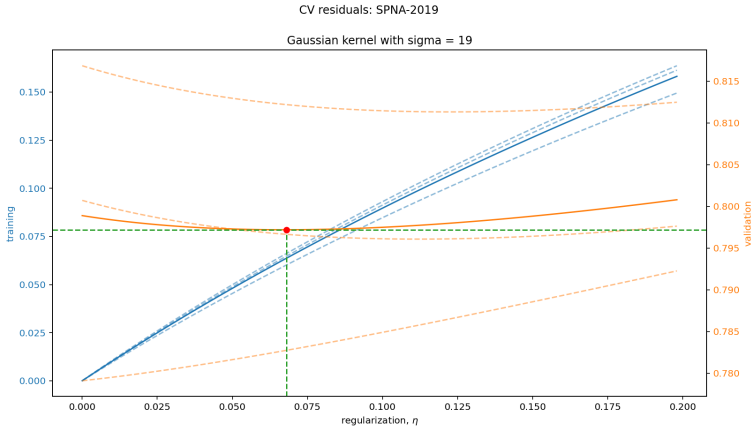



Figure B.5: Residuals of training and validation of the field in SPNA in 2019

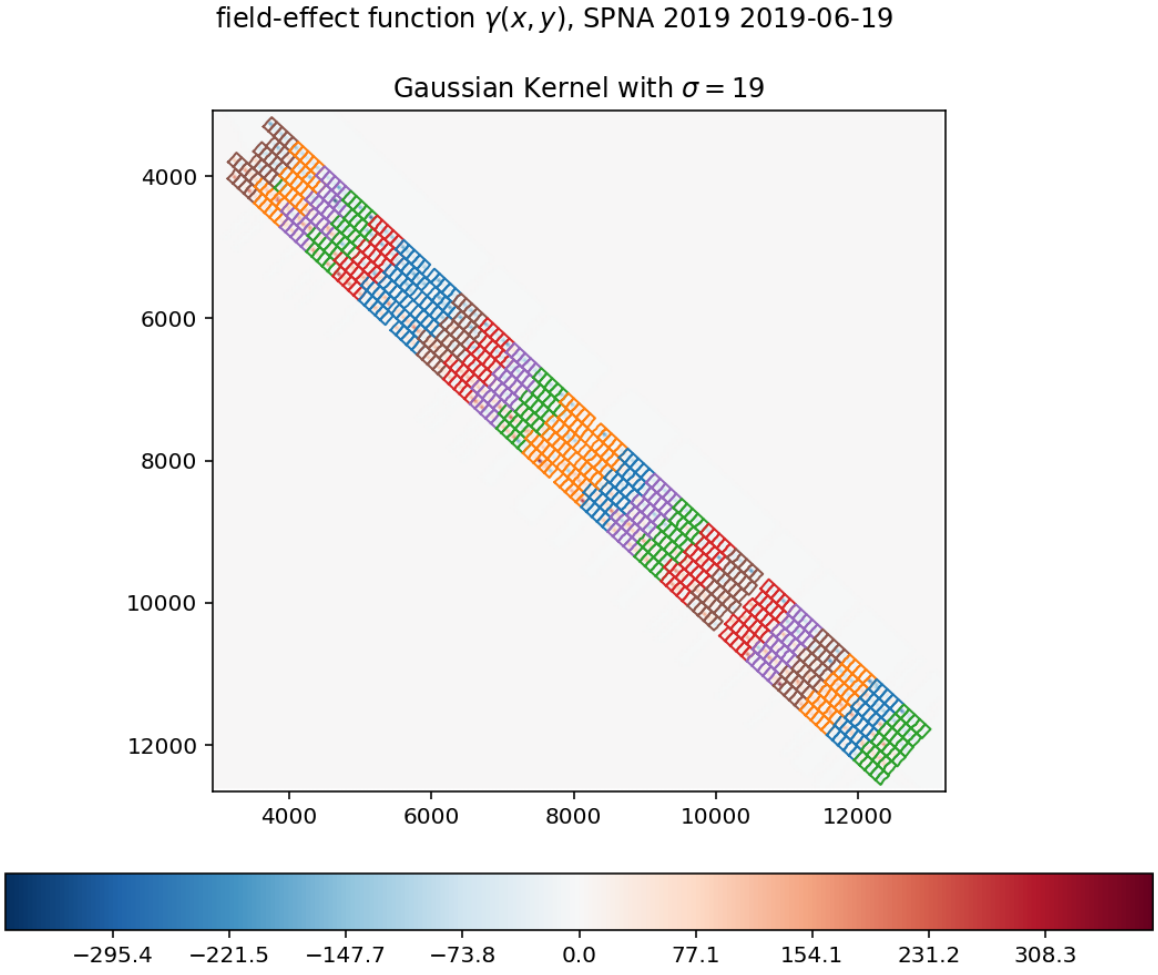
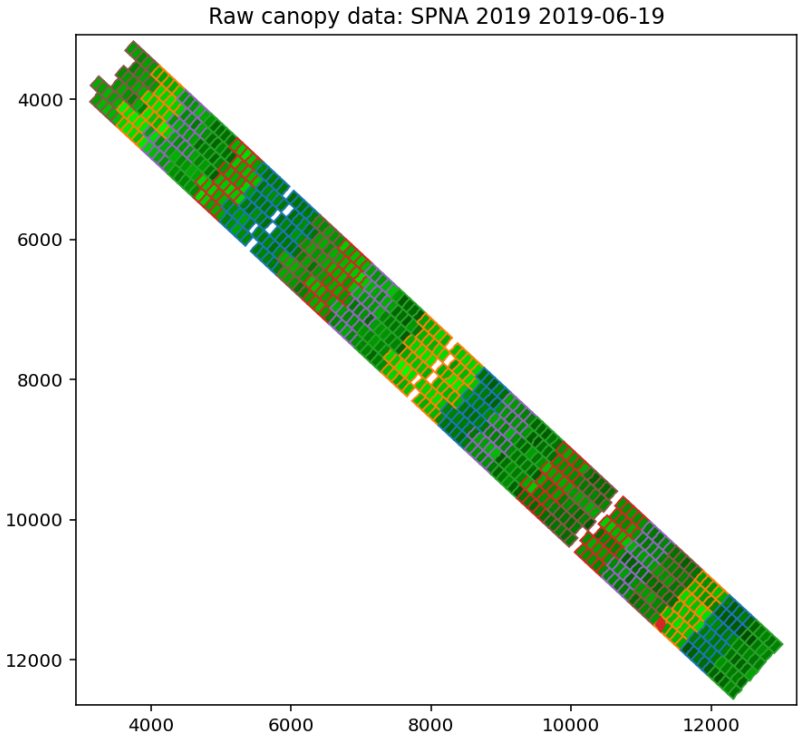
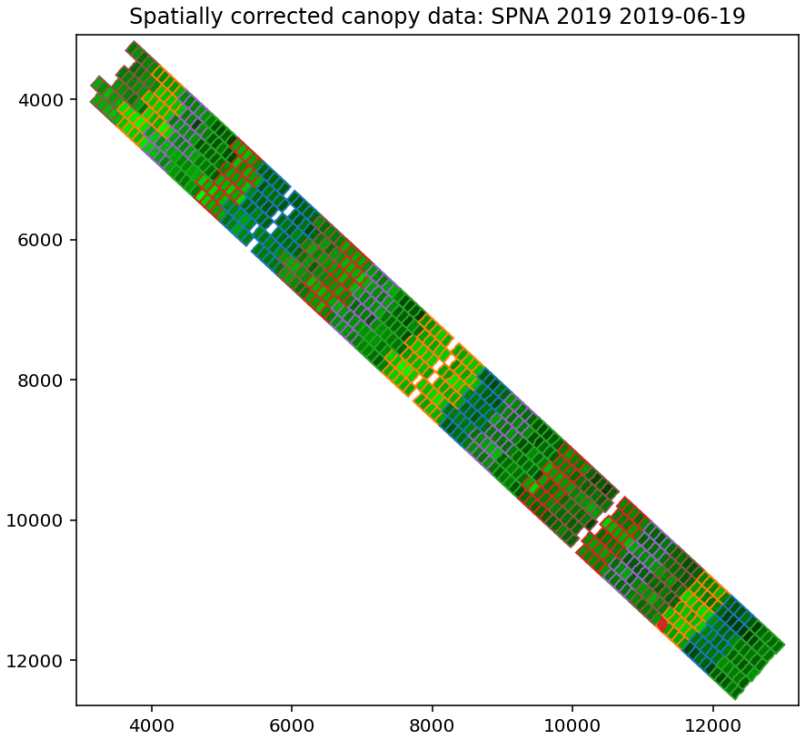



Figure B.6: Field Effect Function on the field in SPNA in 2019




(a)




(b)

Figure B.7: Original (a) and spatial corrected (b) data on the field in SPNA in 2019 with marked target location

## B.4 SPNA 2020

The residuals of SPNA 2020 were included in the main report.

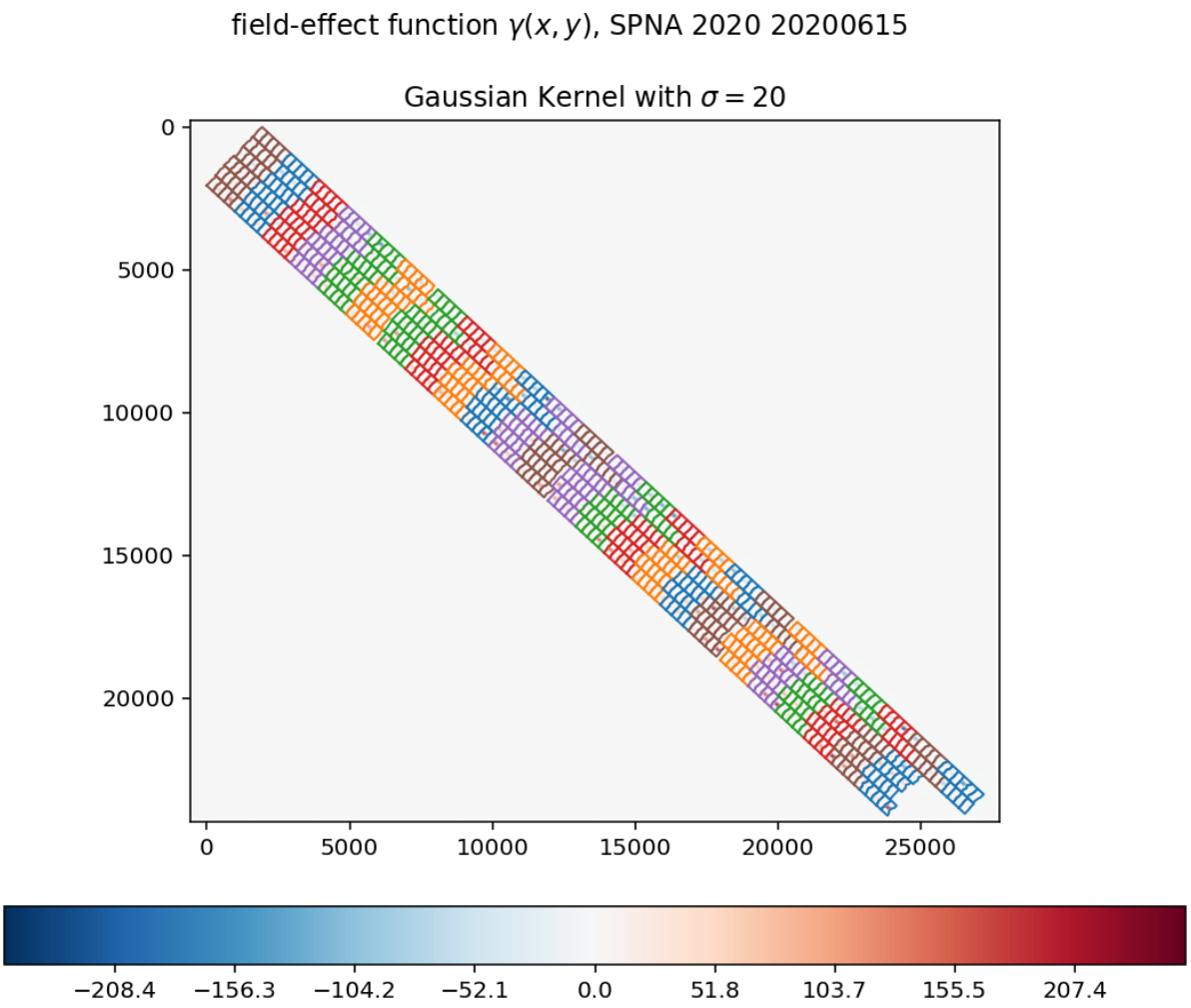



Figure B.8: Field Effect Function on the field in SPNA in 2020

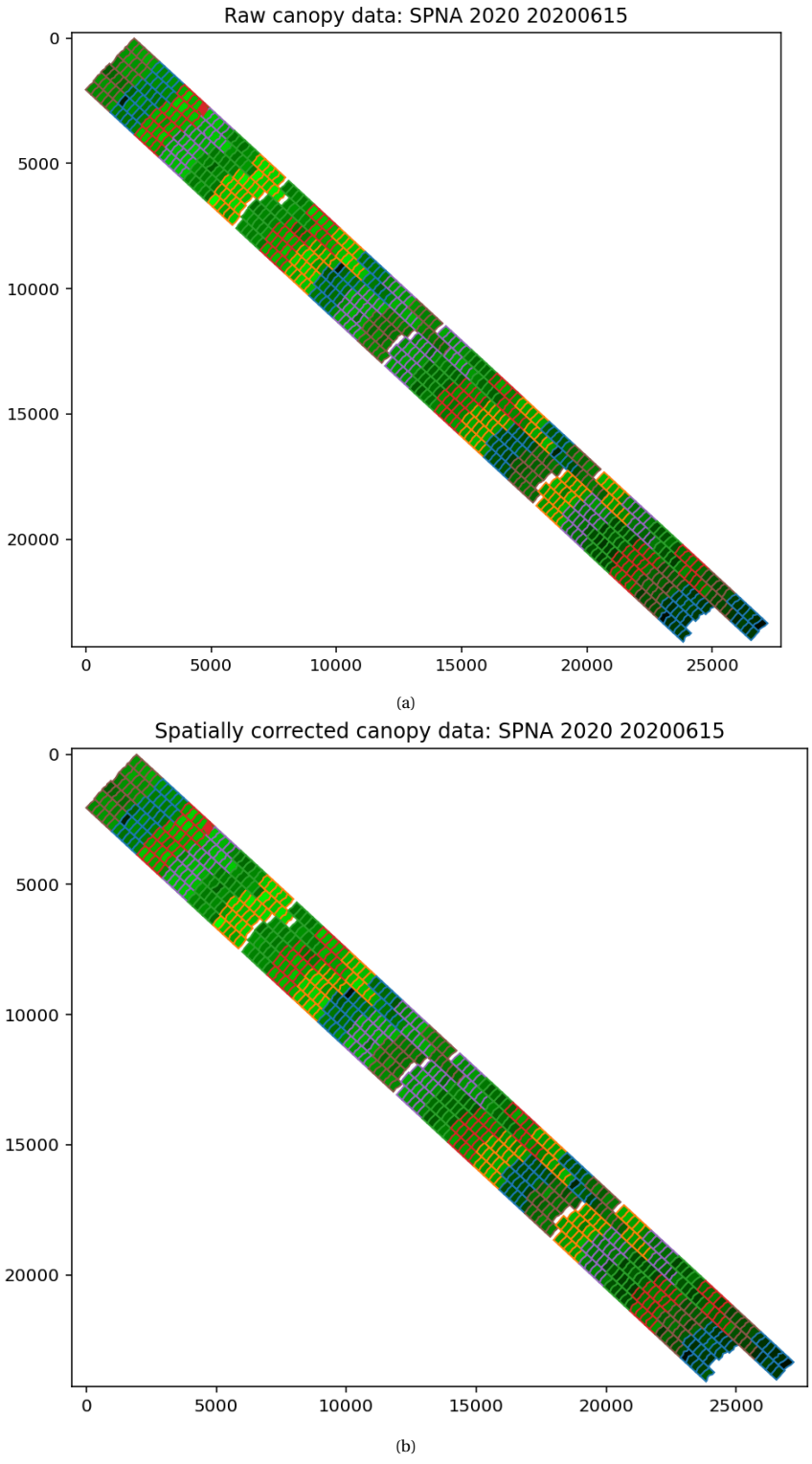



Figure B.9: Original (a) and spatial corrected (b) data on the field in SPNA in 2020 with marked target location

## B.5 SPNA 2021

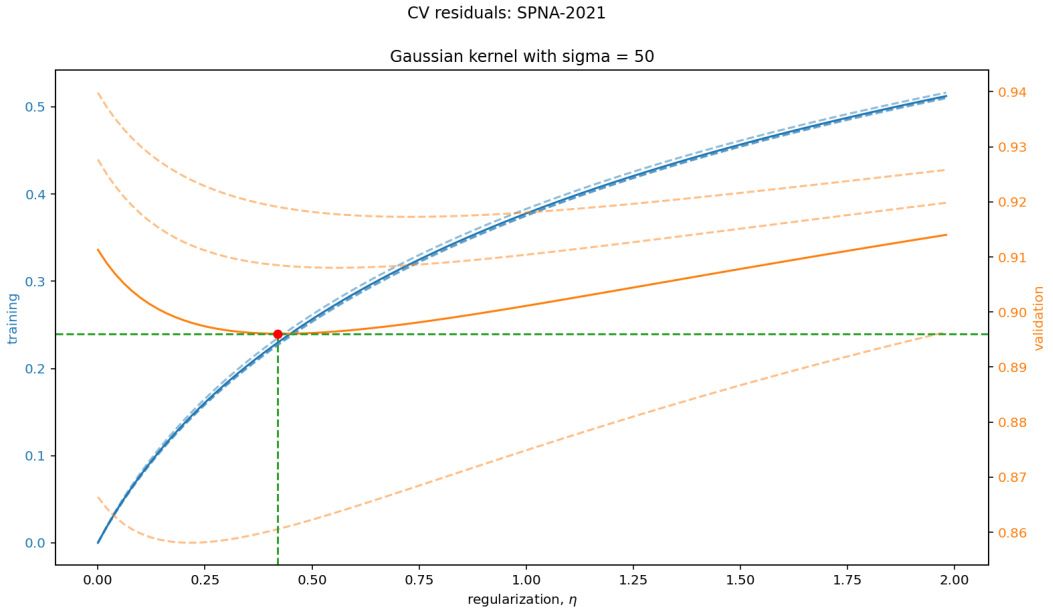



Figure B.10: Residuals of training and validation of the field in SPNA in 2021

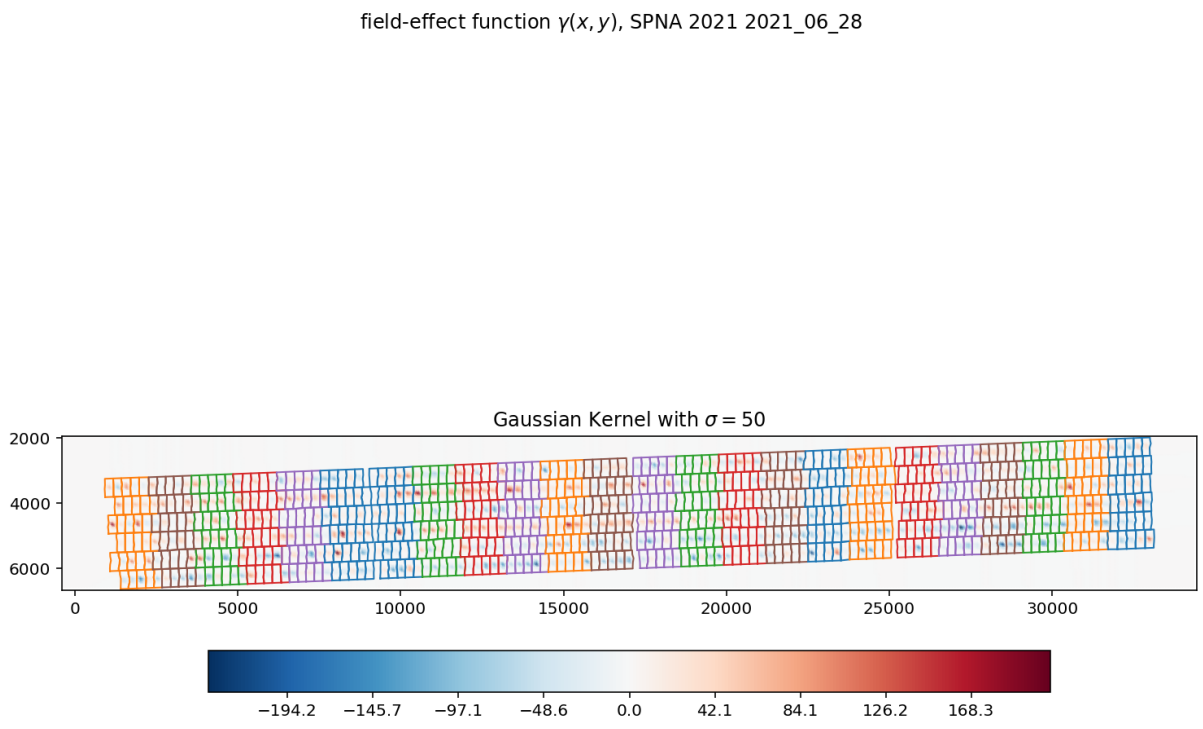
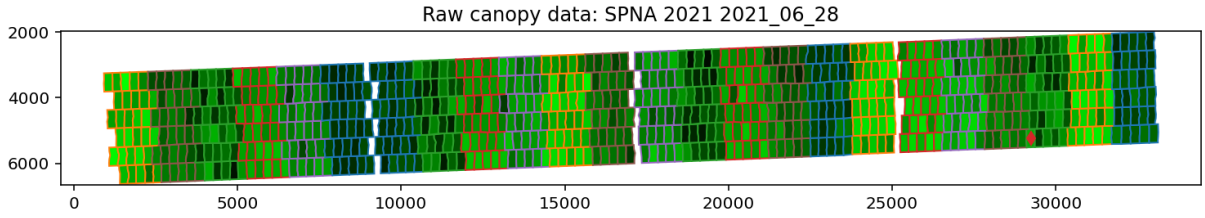
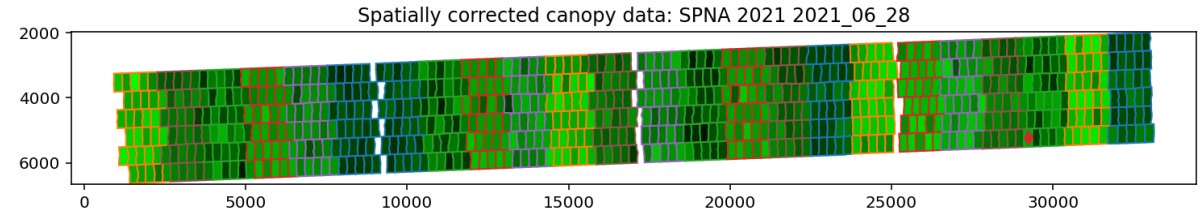



Figure B.11: Field Effect Function on the field in SPNA in 2021




(a)




(b)

Figure B.12: Original (a) and spatial corrected (b) data on the field in SPNA in 2021 with marked target location

## B.6 Veenklooster 2019

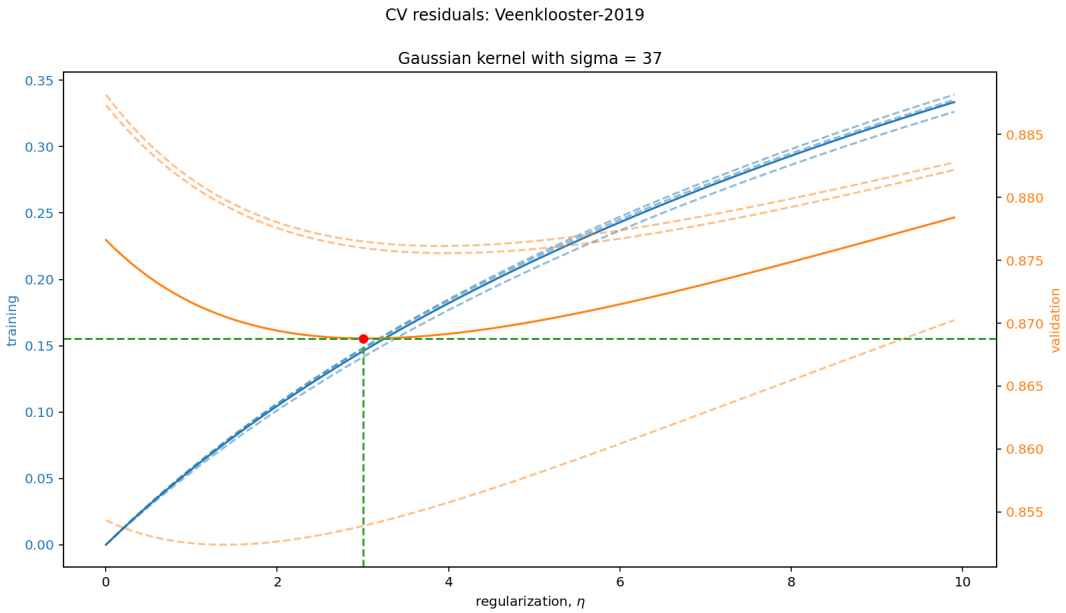



Figure B.13: Residuals of training and validation of the field in Veenklooster in 2019

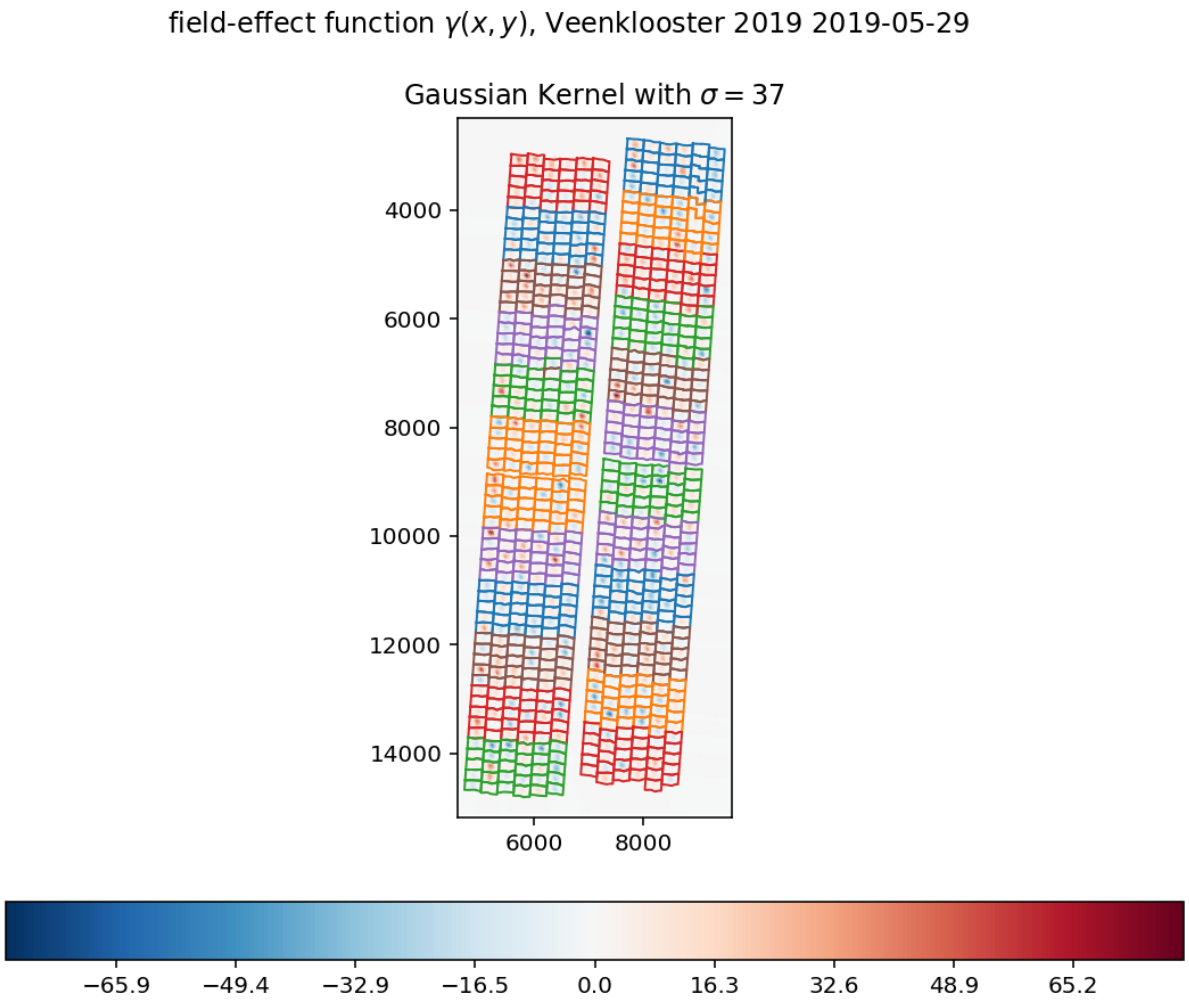



Figure B.14: Field Effect Function on the field in Veenklooster in 2019

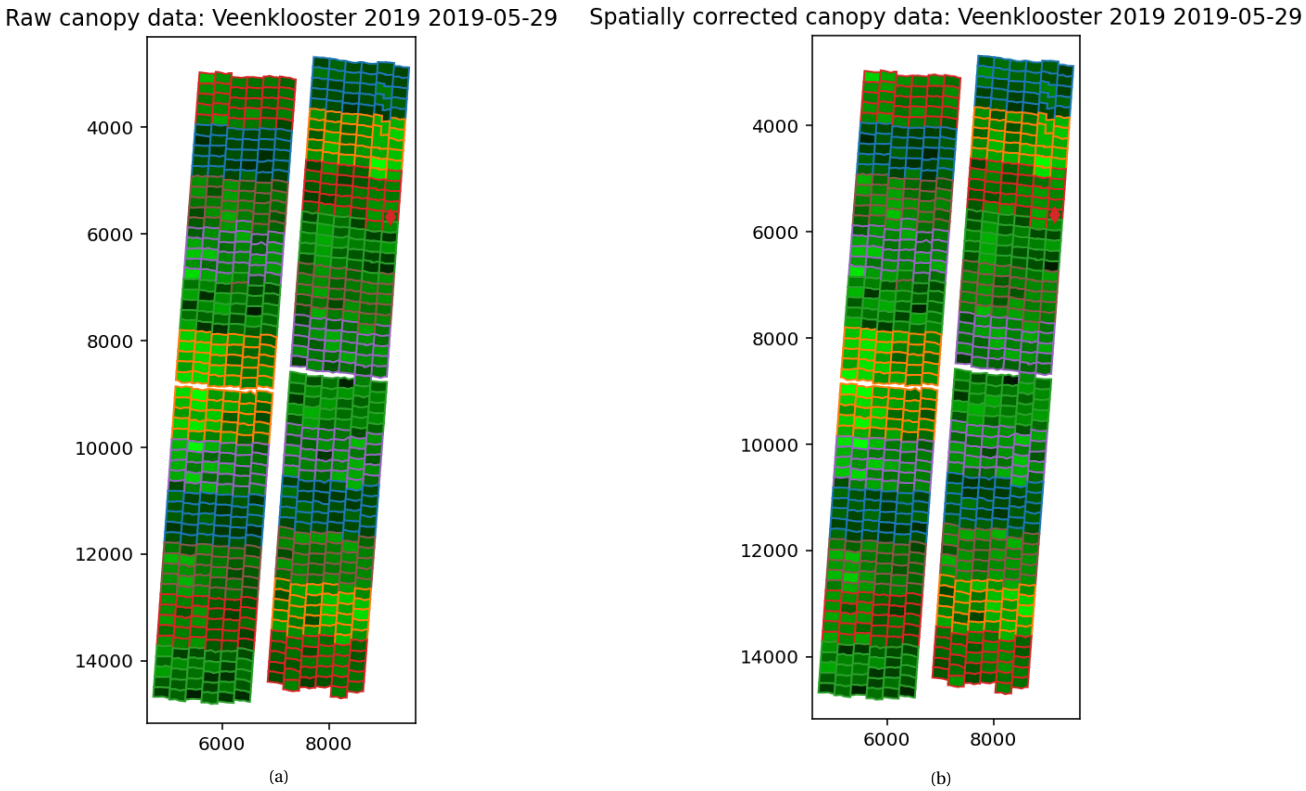



Figure B.15: Original (a) and spatial corrected (b) data on the field in Veenklooster in 2019 with marked target location

## B.7 Veenklooster 2020

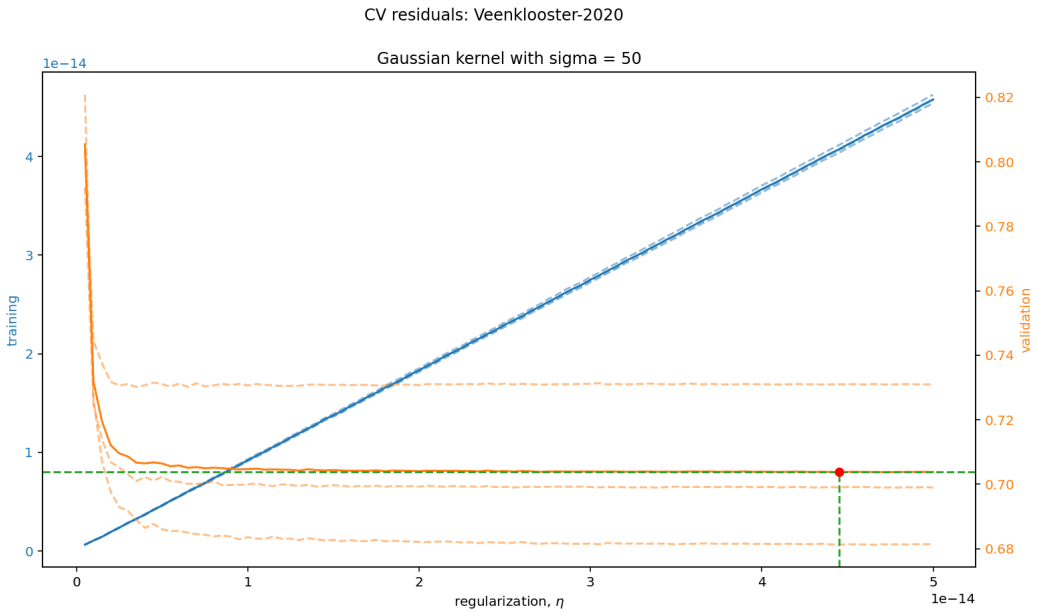



Figure B.16: Residuals of training and validation of the field in Veenklooster in 2020

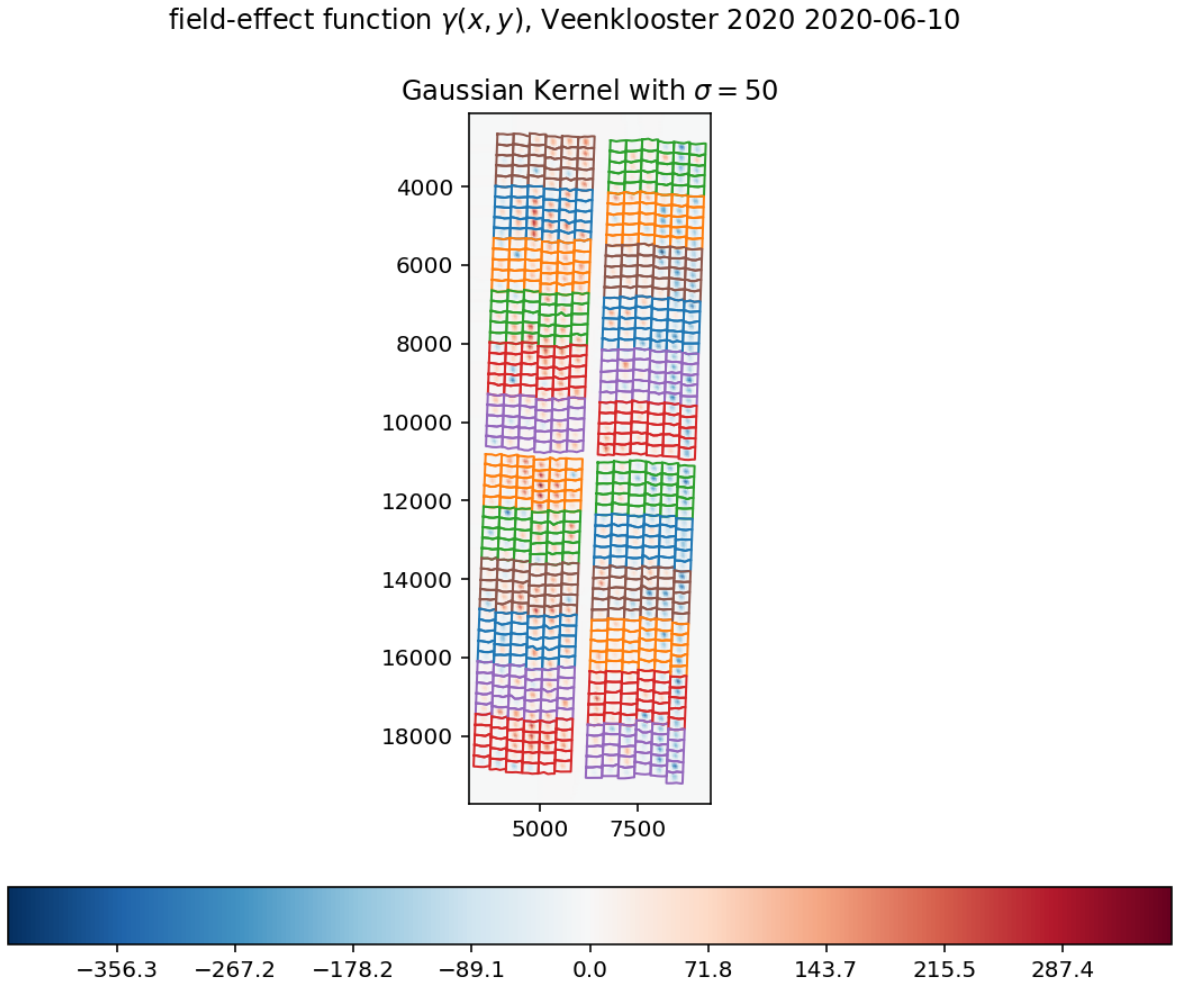



Figure B.17: Field Effect Function on the field in Veenklooster in 2020

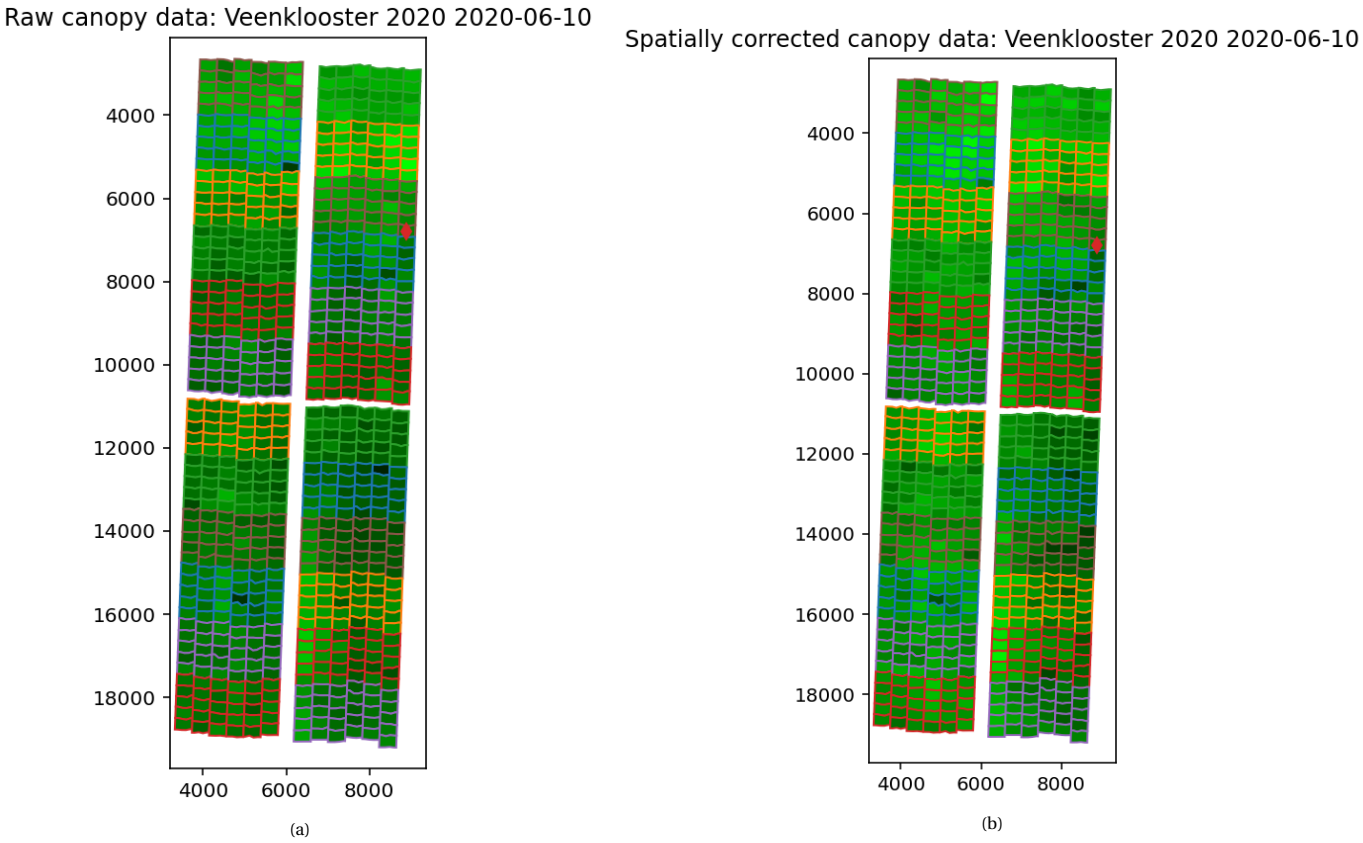



Figure B.18: Original (a) and spatial corrected (b) data on the field in Veenklooster in 2020 with marked target location

## B.8 Veenklooster 2021

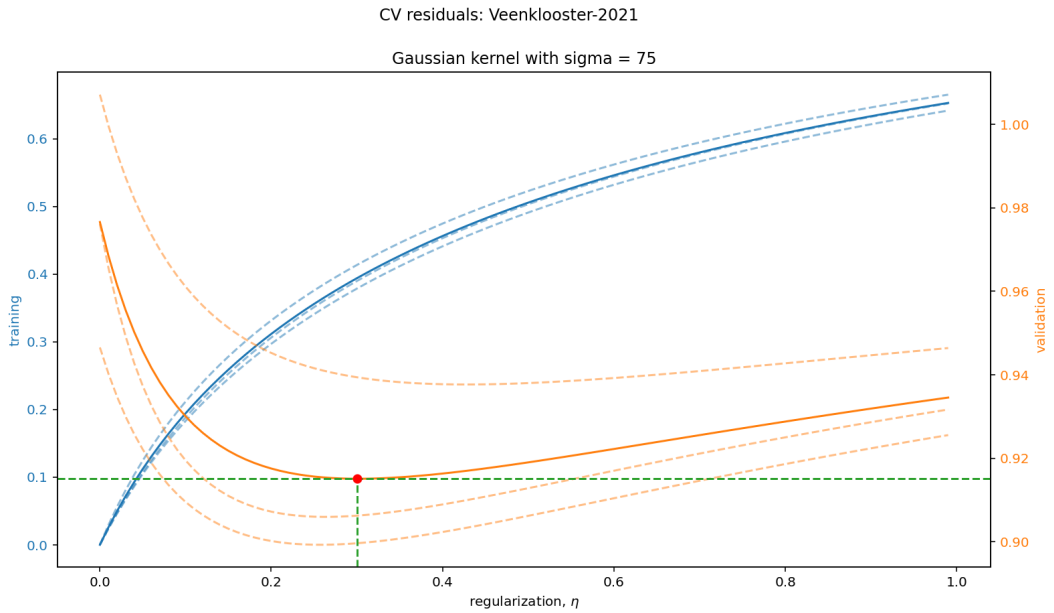



Figure B.19: Residuals of training and validation of the field in Veenklooster in 2021

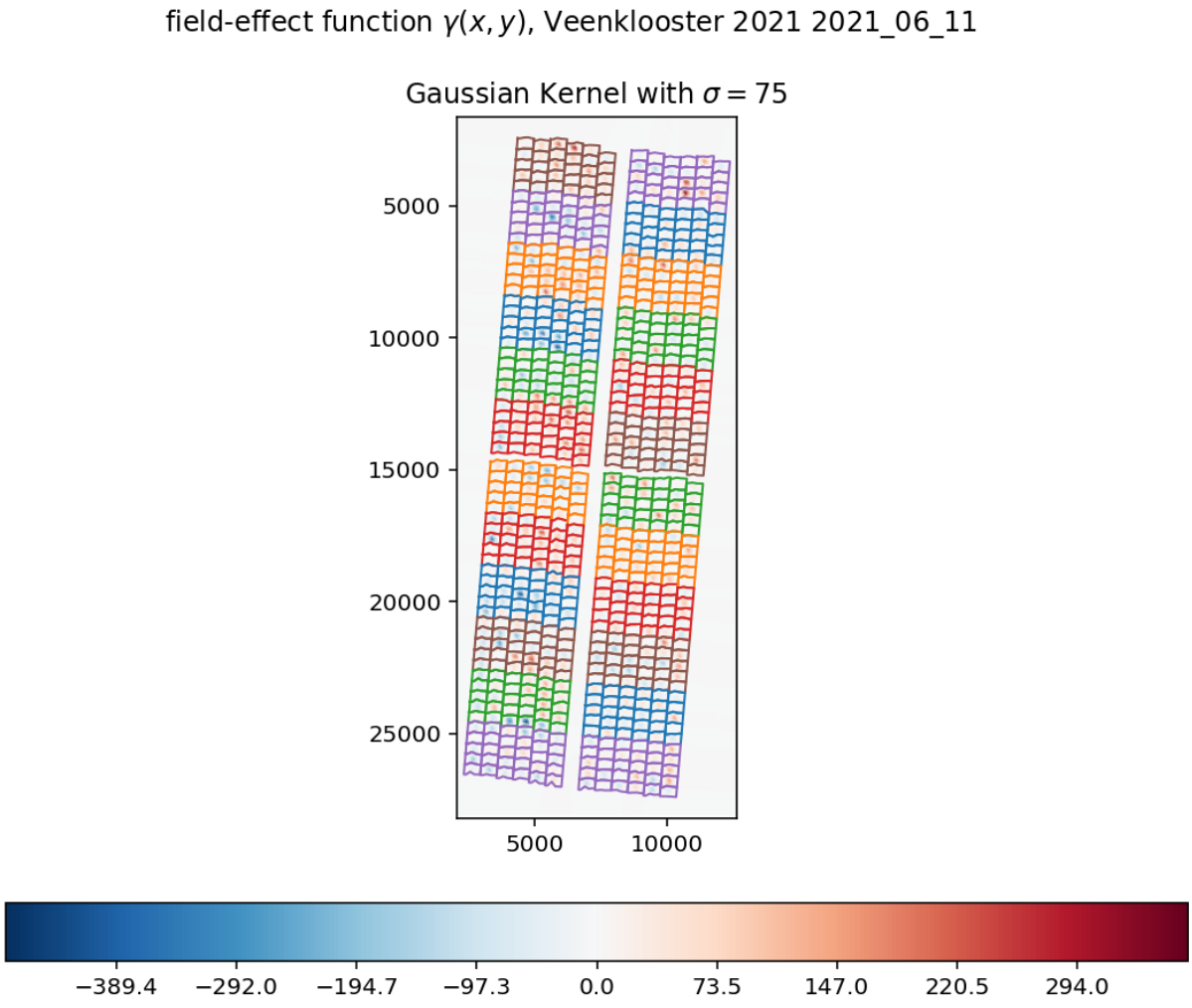



Figure B.20: Field Effect Function on the field in Veenklooster in 2021

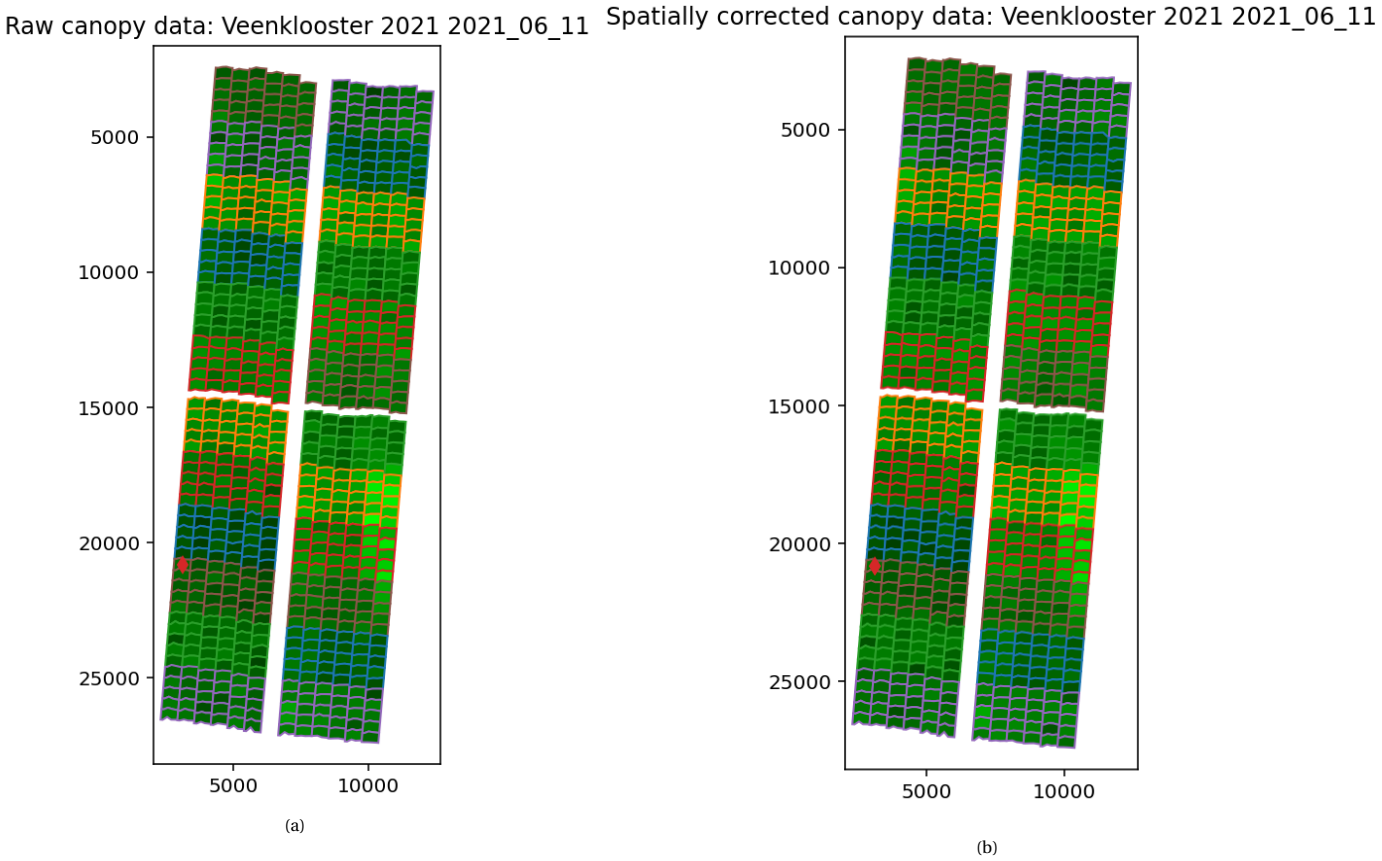



Figure B.21: Original (a) and spatial corrected (b) data on the field in Veenklooster in 2021 with marked target location

# Bibliography


Arfken, G. B., Weber, H. J., & Harris, F. E. (2013). Bessel functions. *Mathematical Methods for Physicists*, 643–713. https://doi.org/10.1016/B978-0-12-384654-9.00014-1

Hofmann, T., Schölkopf, B., & Smola, A. J. (2008). Kernel methods in machine learning. *Annals of Statistics*, *36*, 1171–1220. https://doi.org/10.1214/009053607000000677

Horn, R. A., & Johnson, C. R. (1985). *Matrix Analysis Second Edition*. Cambridge University Press.

Hughes, T. J., Cottrell, J. A., & Bazilevs, Y. (2005). Isogeometric analysis: Cad, finite elements, nurbs, exact geometry and mesh refinement. *Computer Methods in Applied Mechanics and Engineering*, *194*, 4135–4195. https://doi.org/10.1016/j.cma.2004.10.008

Reddy, J. (2006). *An Introduction to the Finite Element Method*. McGraw-Hill Education.

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.

Salvador, F. V., Pereira, G. d. S., Souza, M. H. d., Silva, L. M. B. d., Santana, A. S., de Paula, I. G., Steckling, S. d. M., Fernandes, R. S., Marçal, T. d. S., Carneiro, A. P. S., Carneiro, P. C. S., & Carneiro, J. E. d. S. (2022). Correcting experimental data for spatial trends in a common bean breeding program. *Crop Science*, *62*(2), 825–838. https://doi.org/10.1002/csc2.20703

van Wieringen, W. N. (2015). Lecture notes on ridge regression. *arXiv:1509.09169v7*.