ETH
Swiss Federal Institute of Technology Zurich

Seminar for Statistics

Department of Mathematics

Master Thesis Spring 2022

Lukas Graz

Interpolation and Correction of Multispectral Satellite Image Time Series

Submission Date: September 18th, 2022

Co-Adviser: Gregor Perich
Adviser: Prof. Dr. Nicolai Meinshausen

Preface

Supplementary Material

Instructions and the relevant code needed to reproduce this thesis can be found in the GitHub repository and to use our results we recommend the provided R-package.

More information is given in the appendix **??**.

Acknowledgements

First, I wish to express my sincere gratitude to my supervisor Prof. Dr. Nicolai Meinshausen who took the responsibility for my work and happily took the time to discuss conceptual and guiding questions and to inspire me with new ideas.

This endeavor would not have been possible without Gregor Perich. His high personal commitment, reliability as well as the weekly instructive supervision meetings were essential for this work.

It was a real pleasure for me to be part of the *Crop Science* group for this time. Enjoying everyday company, a two-day excursion, and harvesting wheat together have made this time truly remarkable. In particular, I would like to thank Prof. Dr. Achim Walter, who supported this collaboration at its core.

Last but not least, I would like to express my gratitude to the *Seminar for Statistics*, which created the framework conditions for this work and did everything to help me with conceptional and administrative questions. I should also mention the computing resources provided by them, without which my computations would not have been feasible.

Abstract

Multispectral satellite imagery is used to model vegetation characteristics and development on a large scale in agriculture. As an example, satellite-derived Time Series (TS) of spectral indices like the Normalized Difference Vegetation Index (NDVI) are used to classify crops and to predict crop yield. Sometimes satellite measurements do not match the ground signal due to contamination by atmospheric effects (e.g., clouds or shadows). Therefore, traditional approaches aim to filter out contaminated observations before extracting and subsequently interpolating the NDVI. After filtering, remaining contaminated observations and resulting data gaps are the two challenges for interpolation that we address in this thesis. For this purpose, cereal crop yield maps from 2017-2021 of a farm in Switzerland with the corresponding Sentinel 2 satellite image TS published by the European Space Agency were examined. Contaminated observations were filtered with the provided Scene Classification Layer (SCL). We give a benchmark-supported review of different interpolation methods. Based on it, we found Smoothing Splines as a flexible non-parametric method and Double Logistic approximation as a parametric method with implicit shape assumptions to perform most favorably given the aforementioned challenges. In addition, we generalize an iterative technique which robustifies interpolation methods against outliers by reducing their weights. In most cases, this robustification successfully decreased the 50% and 75% quantiles of the absolute out-of-bag residuals. Moreover, we present a general interpolation procedure that utilizes additional information to correct the target variable with an uncertainty estimate and then performs a weighted interpolation. In our setting, the target variable is the NDVI and as additional information we use the SCL, the observed NDVI and the spectral bands. Consequently, we no longer filter using the SCL, but weight observations according to their reliability. Applying this procedure, the unexplained variance in crop yield estimations via the resulting NDVI TS decreased by 5.4%. Considering the success of the presented procedure with respect to NDVI TS, it appears promising for applications to other satellite-based TS given its cloud-correcting properties.

Contents

List of Figures

List of Tables

Notations

Since this thesis, despite its applied nature, is located at the Mathematics Department, we adhere to the convention of speaking in the first-person plural 'we'.
Furthermore, only equations that are referenced elsewhere are equipped with a number.

Variables

c	A (vector of) constant(s).
$\lambda \in \mathbb{R}$	A scalar.
$n \in \mathbb{N}$	Sample size.
i, j	Indices in $\{1, \ldots, n\}$.
$t \in \mathbb{R}^n$	Time — usually in GDD.
$w \in \mathbb{R}^n$	Vector of weights for each location x.
$y \in \mathbb{R}^n$	Response in 1-dim interpolation setting.
$\hat{y} \in \mathbb{R}^n$	Estimate of y.
$\bar{y} \in \mathbb{R}$	Sample mean of y.
$r \in \mathbb{R}^n$	Residuals given by $y - \hat{y}$.
$X \in \mathbb{R}^{n \times p}$	Design matrix. Each row corresponds to one observation and each column to one covariate.
$X_{[:,j]}$	j^{th} column of X.
$X_{[i,:]}$	i^{th} row of X.

Abbreviations and Objects

DAS	Days After Sowing.
GDD	Growing Degree Days – cumulative sum of "$\max(0, \text{temperature} - \text{threshold})$".
IM	Interpolation Method. That is a method that interpolates data $(t_i, y_i)_{i=1,\ldots,n}$ and yields a function $f(t) = y$, approximating the data.
IS	Interpolation Strategy. By combining various correction models with different IMs (with and without robustification) in section **??** 28 ISs were obtained.
LOOCV	Leave One Out Cross Validation. Describes the procedure of estimating the value for a point by a model that has seen all the points except the current one (see section **??**).
NDVI	Normalized Difference Vegetation Index (**?**).
OOB	Out Of the Box. Describes the procedure of estimating the value for a point by a model that has not seen this point before (see section **??**).
Pixel	A pixel originates of an image pixel and describes a square of 10 x 10 meters in the field that coincides with the resolution (and location) of the Sentinel-2 pixels. Such pixels are illustrated in figure **??**. Additional information like yield is also attached.

P_t	the observed data (weather and spectral bands) at time t and the location of one pixel.
P	a pixel. We see it as a collection of all the observations at the specified location within one season. More formally, $P := \{P_t \mid t \text{ is a valid sample time within a defined season}\}$.
P^{SCL45}	is similar to P but we only consider observations that belong to the classes 4 and 5. This is used done to get a subset of observations which are less contaminated by clouds and shadows.
QAR^{x}	x% Quantile of Absolute Residuals (see definition **??**).
RMSE	Root Mean Square Error (see definition **??**).
S2	Sentinel 2 satellites. Two multi-spectral image satellites deployed by the European Space Agency.
SCL	Scene Classification Layer provided by the European Space Agency that gives an estimation of the land cover class of each pixel. It indicates what one can expect at a pixel at a sampled time. For an overview, see table **??**.
TS	Time Series.
YPE	(Relative) Yield Prediction Error (see definition **??**).

Statistical Models

BS	B-splines (see section **??**).
DL	Double Logistic (see section **??**).
FS	Fourier Series (see section **??**).
GAM	General Additive Model (see section **??**).
LASSO	Least Absolute Shrinkage and Selection Operator (see section **??**).
LOESS	Locally Weighted Regression (see section **??**).
MARS	Multivariate Adaptive Regression Splines (see section **??**).
NW	Nadaraya-Watson (see section **??**).
OLS	Ordinary Least Squares (see section **??**).
OLS^{SCL}	OLS using only the observed NDVI and SCL classes (as factor variables).
OLS^{all}	OLS using the covariates OLS^{SCL} uses and the spectral bands.
RF	Random Forest (see section **??**).
SG	Savitzky-Golay Filter (see section **??**).
SS	Smoothing Splines (see section **??**).
UK	Universal Kriging (see section **??**).

Chapter 1

Introduction

Remote sensing aims to measure target variables efficiently from a distance. Large scale monitoring of forest and agricultural vegetation dynamics is of great interest to authorities, insurance companies and research. Examples include crop classification for subsidizing farmers (**?**) and the creation of crop models for estimating crop yields or nitrogen concentrations (**??**). For this, freely distributed multi-spectral satellite imagery from the Sentinel-2 (S2) satellites are examined (**?**). In order to transform the high dimensional satellite images into easily interpretable metrics, spectral indices such as the Normalized Difference Vegetation Index (NDVI) are used (**?**). The NDVI serves as a proxy for photosynthetic activity (**?**), and thus the corresponding NDVI Time Series (TS) reflects the vegetation development. The quality of a satellite image, however, depends on atmospheric conditions and thus in case of a dense cloud cover, the information content derived from the NDVI is impaired. Therefore, **?** also provides a Scene Classification Layer (SCL), which provides additional metadata about what is observed (e.g., shadows, clouds, vegetation, etc.) . So when extracting the NDVI TS from the Sentinel 2 satellite imagery TS, we can filter out the contaminated observations using the SCL classification. However, due to this filtration it may occur that we have no observations for several weeks, especially in winter. It is also possible that some observations are wrongly classified by the SCL (e.g., as vegetation) and thus result in an erroneous NDVI causing an outlier in the TS. Consequently, the main challenge is to interpolate an NDVI TS, which can contain both large data gaps and outliers.

Currently, there are several approaches to address these issues. One is to look at the observed evolution of the canopy coverage and assume its bell shape for the NDVI TS given the strong correlation between NDVI and photosynthetic activity. Approaches to model this include a 2^{nd} order Fourier approximation (**?**) or a Double Logistic function (**?**). On the other hand, assumptions are made about more abstract properties of the curve, such as smoothness. We divide these into local and global approaches. Nadaraya-Watson (**?**), Savitzky-Golay Filter (**?**) and Locally Reweighed Regression (**?**) use a sliding window to interpolate the TS stepwise. Global methods like B-Splines (**?**) and Smoothing Splines (**?**) reduce the squares of all residuals simultaneously, and Universal Kriging fits a Gaussian process to the data (**?**).

The research questions pursued in this thesis are:

i.) Which IMs are used in the context of NDVI, and what are their advantages and

disadvantages?

ii.) How may contaminated data be dealt with?

iii.) How do data gaps affect interpolation?

iv.) How to deal with data gaps?

v.) How can we recognize a good interpolation of the NDVI?

In this thesis, we will discuss the strengths and weaknesses of Interpolation Methods (IMs) and evaluate them with respect to NDVI interpolation. For this purpose, we use the Sentinel 2 satellite image TS and crop yield maps of different fields of different cereal species on a farm in Witzwil, Switzerland over the years 2017-2021. After presenting the available data, illustrating challenges and defining different concepts in chapter **??** (**??**), we turn to the two main blocks of this thesis. One covers the study of IMs and the other presents a general procedure of correcting (NDVI) TS with uncertainty estimation by utilizing additional information. On the first block, in chapter **??** (**??**) we examine parametric and non-parametric IMs and discuss their strengths and weaknesses (question i.). We generalize and test an iterative technique that makes IMs more robust to outliers by weighting them less (question ii.). To evaluate IMs, we present an approach that uses out-of-bag residuals (question v.). In section **??** (**??**), we discuss how different IMs respond to data gaps (question iii.), and in section **??** (**??**) we preselect IMs. We evaluate this preselection in the results section **??** (**??**) and select two candidates from different IMs in section **??** (**??**). For the second block, we correct possibly contaminated data with statistical models in chapter **??** (**??**) (question ii.) and utilize previously ignored observations, which we hope will further reduce data gaps (question iv.). Thus, we no longer filter the observations a priori via the SCL, but instead correct the observed NDVI and weight the observations via estimated uncertainties. By combining different statistical models and IMs, we get 28 Interpolation Strategies (ISs). We compare those with a vegetation-oriented quality measure (question v.) and describe the results in section **??** (**??**). Based on these results, in section **??** (**??**) we argue what the best IS is. In addition, we justify why our NDVI correction can be understood as unsupervised learning and why we relied only on satellite imagery and not on meteorological data for the NDVI correction. Our conclusions of this thesis, recommendations, as well as an outlook on future work is given in chapter **??** (**??**).

Chapter 2

Data and Methods

This section describes the available data and the challenges associated with it. Our study region is a farm of over 800ha, which is located in western Switzerland. From **?** we acquired Sentinel-2 (S2) satellite image data (section **??**), yield maps of several cereals from 2017 to 2021 (section **??**), and meteorological data (section **??**). Methods to evaluate an estimator or model are given in section **??**.

For IMs we refer to sections **??** and **??** for a robust interpolation technique to section **??**. In section **??** we describe a method to objectively determine the quality of an interpolation, and in chapter **??** we present a strategy which involves correction of the NDVI with a weighted interpolation.

2.1 Sentinel 2 Data

The European Space Agency (**?**) freely distributes images of the S2 satellites. Together, both satellites have a revisit time of 5 days in our study region.

The S2 images contain 12 spectral bands with spatial resolutions of up to 10 meters (see table **??**). Bands with a lower resolution (20 and 60 meters) were upscaled to 10 meter resolution using cubic interpolation (**?**). In order to decrease the effect of atmospheric conditions like reflections and scattering, bottom-of-atmosphere, radiometric corrected Level-2A data was used. **?** supplies the Scene Classification Layer (SCL). It is a model output that for each location assigns the observed pixel to one of 11 SCL-classes (cf. table **??**). In this thesis, we will use this classification to filter out data points that we believe to be less informative. These are all observations in which the SCL-class does not correspond to vegetation or bare soils (classes 4 and 5). We define the set SCL45 as the observations that are labelled as SCL-class 4 or 5.

2.2 Crop Yield Data

The crop yield data were collected from a combine harvester. Equipped with a satellite-based navigation system, the harvester drives over the fields and continuously estimates the dry crop yield density in t/ha (see figure **??**). We use the data set presented in **?**, where error-prone measurement points (such as during a tight curve of the combine harvester) were removed and then the yield map was rasterized using linear interpolation (cf. figure **??**). We summarize the rasterized dry yield values by the following statistics:

Table 2.1: List of spectral bands of the S2-satellites. Each band has its center at the wavelength λ in nm with the spectral width $\Delta\lambda$ in nm with a spatial resolution SR in m (**?**).

Band	λ	$\Delta\lambda$	SR	Purpose
1	443	20	60	Atmospheric correction (aerosol scattering)
2	490	65	10	Sensitive to vegetation senescing, carotenoid, browning and soil background; atmospheric correction (aerosol scattering)
3	560	35	10	Green peak, sensitive to total chlorophyll in vegetation
4	665	30	10	Maximum chlorophyll absorption
5	705	15	20	Position of red edge; consolidation of atmospheric corrections / fluorescence baseline.
6	740	15	20	Position of red edge, atmospheric correction, retrieval of aerosol load.
7	783	20	20	Leaf Area Index (LAI), edge of the Near-Infrared (NIR) plateau.
8	842	115	10	LAI
8a	865	20	20	NIR plateau, sensitive to total chlorophyll, biomass, LAI and protein; water vapor absorption reference; retrieval of aerosol load and type.
9	945	20	60	Water vapor absorption, atmospheric correction.
10	1375	30	60	Detection of thin cirrus for atmospheric correction.
11	1610	90	20	Sensitive to lignin, starch and forest above ground biomass. Snow/ice/-cloud separation.
12	2190	180	20	Assessment of Mediterranean vegetation conditions. Distinction of clay soils for the monitoring of soil erosion. Distinction between live biomass, dead biomass and soil, e.g., for burn scars mapping.

Table 2.2: Overview: Scene Classification Layers (SCL)

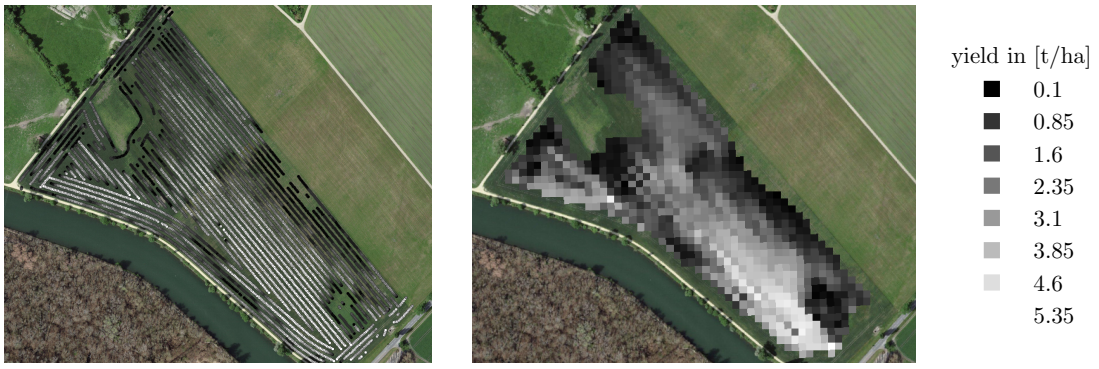
Color	No.	Class	Color	No.	Class
	0:	Missing Data		6:	Water
	1:	Saturated or defective pixel		7:	Cloud low probability
	2:	Dark features / Shadows		8:	Cloud medium probability
	3:	Cloud shadows		9:	Cloud high probability
	4:	Vegetation		10:	Thin cirrus cloud
	5:	Bare soils		11:	Snow or ice

Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum	Variance
0.107	6.186	7.560	7.359	8.756	13.35	4.035

Comparing the average per-field crop yield reported by the farmer with the yield estimated by the combine harvester shows that the latter overestimates crop yield by ca. 10% (**?**). Since the relative estimation error is approximately constant and we do not aim for an absolute yield prediction, we will not consider this deviation.

2.3 Normalized Difference Vegetation Index

The well-known NDVI introduced by **?** is used to approximate vegetation in remote sensing. It utilizes the fact that healthy, photosynthesizing vegetation exhibits a large increase in reflectance between the red and infrared region of the light spectrum (cf. appendix figure **??**). It is calculated using the S2 bands $B4$ (red) and $B8$ (infrared)



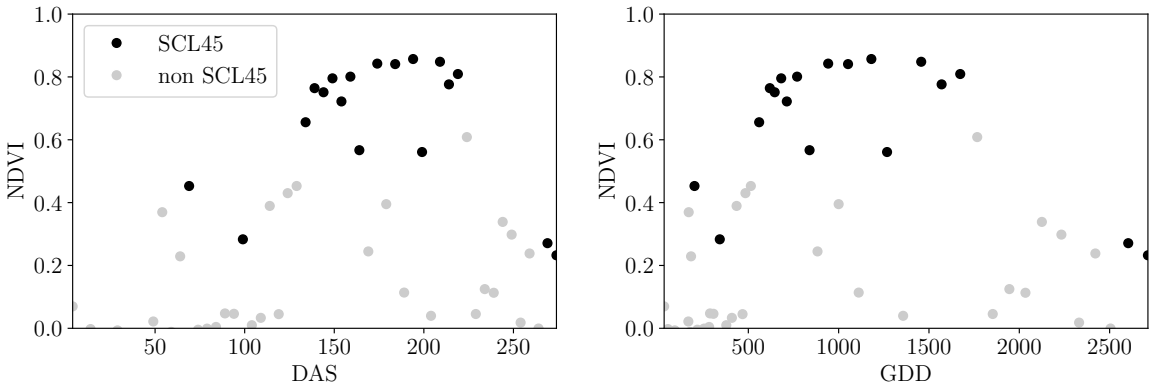
(a) Cleaned combine harvester data (b) Rasterized to Sentinel 2 resolution.

Figure 2.1: Crop yield density map of a field.

(table **??**) by:

$$NDVI = \frac{B8 - B4}{B8 + B4}$$

Since we measure the NDVI via the S2 satellites from space, we cannot expect to obtain the same NDVI as measured with a ground-based spectroradiometer. This is especially true if the ground signal is obfuscated by either the clouds directly or by cloud shadows. Even if we only use SCL45 observations flagged as cloud-free, we still encounter measurement errors, as described in section **??**. Therefore, we call the calculated values merely the observed NDVI. In the following chapters, we will study the resulting NDVI TS extensively. Such a TS is shown in figure **??**.



(a) Days After Sowing (DAS) (b) Growing Degree Days (GDD)

Figure 2.2: NDVI TS plotted against DAS and GDD. GDD are introduced in section **??**.

2.4 Transformation of Timescale

Two drawbacks become apparent when using Days After Sowing (DAS) as the timescale (see figure **??**): First, this scale makes it difficult to compare two NDVI TS because wheat is not always sown on the same day of the year and in some years plants begin to emerge earlier. Second, because there are only few SCL45 observations in autumn and winter, we face significant data gaps during this period. To fix both problems, **?** propose to transform the timescale into a meaningful temperature based one. The resulting Growing

Degree Days (GDD) are defined as the cumulative sum of temperature above a given base temperature T_{base} since sowing. For cereals, we use $T_{base} = 0$ (**?**). Thus, the GGD for n days after sowing will be equal to:

$$GDD_n := \sum_{i=0}^{n} \max(T_i - T_{base}, 0).$$

Important plant growth stages and their corresponding GDD values are tabulated in the appendix **??**.

In figure **??** we see an example for comparison of the DAS and GDD timescale. Here we see that the first 120 DAS are compressed to just 500 GDD and hence the gap in observations was successfully compressed. Given the reasons mentioned above, from now on we will only consider GDD.

2.5 The Concept of a 'Pixel'

Now we create a new data structure that we call Pixel. This originates from the pixels of the S2 satellite images. It will contain all the information needed to answer the research questions in the following chapters.

Consider a 10 by 10 meter square that coincides with a S2 image pixel and T the GDD values for which S2 images are available in a given season. For $t \in T$ let P_t be a tuple of all the spectral bands, the observed NDVI and the SCL class at the considered location at time t. Then, define P as the collection of all the P_t and the estimated dry yield for this square. Analogously to P, define P^{SCL45} by only considering P_t with SCL-classes 4 or 5 (vegetation and soil).

2.6 Illustration of S2 Images

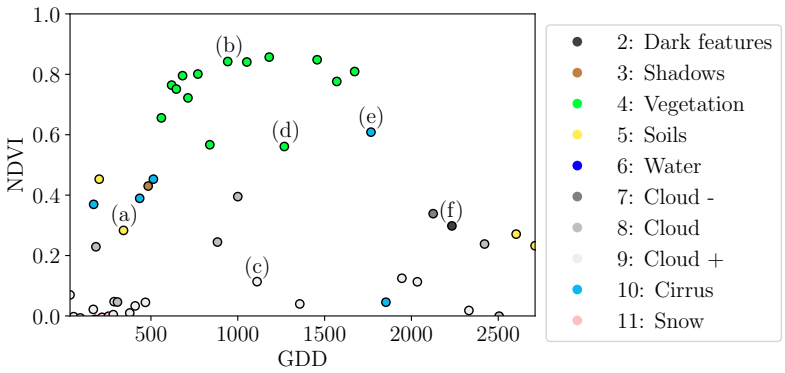
Using an example pixel, we illustrate the challenges in working with S2 image data. Figure **??** shows a selection of 6 satellite images of a field, one selected Pixel and the NDVI TS of this pixel. In February (image a), we see no vegetation but bare soil and thus also a low NDVI. At the beginning of May (b), we observe a cloudless dark green field with a high NDVI. In (c) heavy cloud cover (SCL class 9) leads to a complete loss of plant information in this S2 observation. Figure (d) shows that the SCL classification is not reliable, since we evidently observe clouds, which is also reflected in a sudden NDVI drop. Even though SCL indicates that (e) are thin cirrus clouds, a pale green shimmers through, and we also note a reasonably high NDVI. Therefore, we remark that some SCL45 observations are not accurate and even though a few non-SCL45 observations contain useful information, most of them are too unreliable (e.g., all SCL 9 observations). Thus, we aim to substitute the unreliable ones with interpolated versions and correct contaminated ones.

2.7 Estimation Evaluation Criteria

In this section, we define score functions for estimation accuracy and techniques to adequately apply them.

(a) 2021-02-23 SCL 5 (b) 2021-05-09 SCL 4 (c) 2021-05-24 SCL 9

(d) 2021-06-03 SCL 4 (e) 2021-06-28 SCL 10 (f) 2021-07-23 SCL 2



(g) Corresponding NDVI TS

Figure 2.3: Satellite images of a field at selected times with a static greyscale background for orientation. Moreover, the NDVI TS of the red-highlighted pixel is shown in (g) colored by the SCL labels.

2.7.1 Score Functions

Now, we define the Root Mean Square Error (RMSE), a prominent and outlier-sensitive score function.

Definition 2.7.1.1. *(RMSE) Given a vector $y \in \mathbb{R}^n$ and its estimator $\hat{y}$, we define the RMSE as:*

$$\text{RMSE} := \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Thus, the RMSE measures how close the estimated $\hat{y}$ are to the original y by considering the squares of errors. The lower the RMSE, the more similar are the fitted values to the original values. Note that one strong outlier may corrupt this score function. Therefore, we will also define the x^{th}% Quantile of Absolute Residuals (QAR^{x}) that can handle some fraction of outliers:

Definition 2.7.1.2. *(QAR^x) Given a percentage $x \in \{1, 2, \ldots, 100\}$, a vector $y \in \mathbb{R}^n$ and its estimator $\hat{y}$, we assume that $|y_1 - \hat{y}_1| \leq |y_2 - \hat{y}_2| \leq \cdots \leq |y_n - \hat{y}_n|$. Then, we define the QAR^x as the biggest residual under the x% smallest ones. Formally:*

$$\text{QAR}^x := \max\left\{|y_i - \hat{y}_i| : i \leq \frac{x}{100}n\right\}$$

Note that QAR^{50} coincides with the median of absolute residuals and QAR^{100} with the maximum of absolute residuals. Hence, the higher, x the fewer outliers QAR^{x} can handle. Consequently, if we expect the data to have 5% outliers, we should choose x smaller than 95. Furthermore, if an estimator attains lower QAR^{x} values than a competing estimator for $x = 50, 75, 85$, we conclude[1] that it also produces smaller residuals in most cases.

2.7.2 Out-Of-Bag and Leave-One-Out-Cross-Validation

The rationale for Out-Of-Bag (OOB) and Leave-One-Out-Cross-Validation (LOOCV) is that we intend to evaluate a model M with unseen data. That is, if D describes the entire dataset, and we train a model on a subset of D, we can use the remaining data to evaluate the model. To formally introduce this, let:

$$D = \{(X_{[j,:]}, y_j) |\ X \in \mathbb{R}^{n\times p}, y \in \mathbb{R}^n, j = 1, \ldots, n\}$$

be a dataset, $i \in \{1, \ldots, n\}$ and $M^{(-i)}$ a model fitted on a subset of $D \setminus \{(X_{[i,:]}, y_i)\}$. Then we call $\hat{y}_i := M^{(-i)}(X_{[i,:]})$ an OOB estimator of y_i. If we do this for all $i \in \{1, \ldots, n\}$, we obtain $\hat{y} := (\hat{y}_1, \ldots, \hat{y}_n)$ the OOB estimator for $y \in \mathbb{R}^n$.

In the case that $M^{(-i)}$ was fitted on the set $D \setminus \{(X_i, y_i)\}$ (i.e., not a true subset), we call the corresponding $\hat{y}_i$ also the LOOCV estimator. If we optimize some parameter via OOB (or LOOCV) this means that we search for the parameter that minimizes some loss function which takes the OOB (or LOOCV) residuals. In the bootstrap (e.g., random forest) framework, we define $\hat{y}_i$ to be the average of all computed and admissible $M^{(-i)}$.

[1] Strictly speaking, this conclusion assumes that the distribution of the residuals is unimodal (i.e., has only one 'peak').

Chapter 3

Interpolation

The need for interpolating the NDVI TS was established in the precious chpater. In this chapter, we first specify a setting for the interpolation and categorize the IMs into those that make fundamental shape assumptions (parametric) and those that are more flexible (non-parametric). We give an introduction for each method with a compact definition, highlight adjustments or give remarks where appropriate, and point out strengths and weaknesses of each method. A brief overview of the considered IMs is provided in table **??**. Afterwards, we extract a robustification strategy from one IM and generalize it, so we can use it for all methods that allow for a priori weighted observations. Finally, using LOOCV, we tune the parameters (where necessary) and get a first idea of the performance of each method.

3.1 Interpolation Setup

For now, we will only consider SCL45 observations since they are more reliable. Hence, data in the form of (t_i, y_i) for $i = 1, \dots, n$ is given, where t_i is the time in GDD and y_i denotes the NDVI at t_i. We assume that it can be represented by

$$y_i = m(t_i) + \varepsilon_i,$$

where ε_i is some random noise and $m : \mathbb{R} \to \mathbb{R}$ is some (parametric or non-parametric) function. If we assume that $\varepsilon_1, \dots, \varepsilon_n$ i.i.d. with $\mathbb{E}[\varepsilon_i] = 0$ then

$$m(t) = \mathbb{E}[y \mid t]$$

We will introduce parametric and non-parametric approaches to estimate m in section **??** and **??**. Furthermore, in the subsequent, we denote $w \in \mathbb{R}^n$ as the vector of weights such that w_i corresponds to the weight that (t_i, y_i) should have in the interpolation.

3.2 Parametric Regression

Parametric Curve estimation tries to fit a parametric function, such as, for example, a Gaussian function with parameters μ and σ, to a dataset. In the following, we introduce two parametric approaches.

3.2.1 Double Logistic

The Double Logistic (DL) smoothing as described in **?** heavily relies on shape assumptions of the fitted curve (i.e., the NDVI TS). First, we assume that there is a minimum NDVI level $y_{\min}$ in the winter (e.g., due to evergreen plants), which might be masked by snow. This can be estimated beforehand, taking several years into account. Second, we assume that the growth cycle can be divided into an increase and a decrease period, where the TS follows a logistic function. The maximum increase (or decrease) is observed at t_0 (or t_1) with a slope of d_0 (or d_1). The equation of the double-logistic fit is given by:

$$y(t) = y_{\min} + (y_{\max} - y_{\min}) \left(\frac{1}{1 + e^{-d_0(t-t_0)}} + \frac{1}{1 + e^{-d_1(t-t_1)}} - 1 \right)$$

Where the five free parameters: $y_{\max}$, d_0, d_1, t_0, t_1 are initially estimated by least squares. Such fit can be seen in figure **??**.

Robustification

Similar as for the SG (introduced in section **??**) one can reestimate the parameters by giving less weight to the overestimated observations and more weight to the underestimated observations. For the details on the choice of the weights, we refer to **?**. We will not apply this reestimation, but rather the robustification introduced later in section **??**.

Advantages	Disadvantages
— Incorporates subject specific knowledge in the case of evergreen plants covered in snow.	— Strong shape assumptions limit flexibility.
— Optimized parameters have an intuitive meaning.	— Parameter optimization might go wrong. This can be mitigated to some extent by providing bounds for the parameters.
— Robust due to shape assumptions.	
— Meaningful parameters.	

3.2.2 Fourier Series

? approximates the NDVI curve using a second order Fourier Series (FS):

$$\mathrm{NDVI}(t) = \sum_{j=0}^{2} a_j \times \cos(j \times \Phi_t) + b_j \times \sin(j \times \Phi_t)$$

where $\Phi = 2\pi \times (t-1)/n$. If we set the period to match one year, this would coincide with the notion that plants grow every year. Analogous to section **??** we fit it to the data by least squares. Example fits can be seen in figure **??**

Advantages	Disadvantages
— Assumption of periodicity can be helpful if we are modelling multiyear grow cycles.	— Bad behavior in regions with little data (cf. figure **??**).
— Fourier Series are widley known.	— Hard to interpret estimated parameters.
	— Parameter estimation can go wrong. Introducing bounds can help.

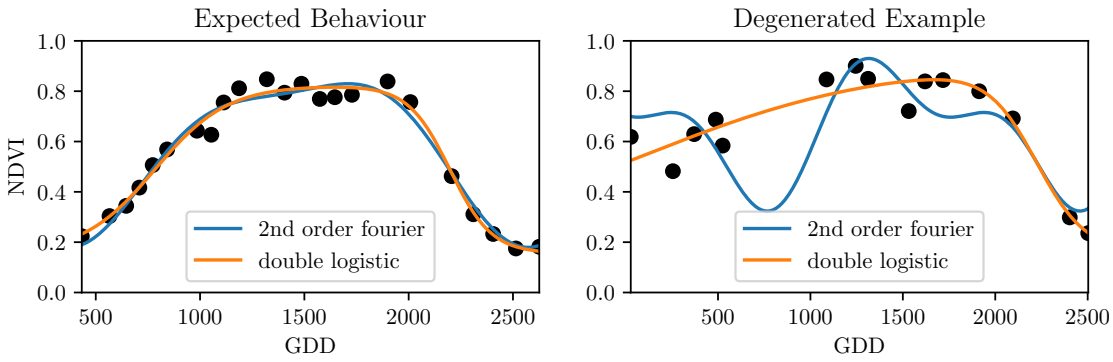


Figure 3.1: Here we observe the possibilities of a precise fit for the two parametric methods but notice also some misbehavior.

3.2.3 Optimization Issues

We shall mention some optimization issues we countered during implementation. Since we aim to minimize the residual sum of squares over 5 (or 6) parameters, we try to solve a non-convex optimization problem. Thus, the algorithm[1] either struggles to find the global minimum or fails to converge. This was fixed by providing for each parameter reasonable initial values and generous bounds (that match our experience).

3.3 Non-Parametric Regression

In non-parametric curve estimation, the curve does no longer have to be fully determined by parameters, but we allow it to flexibly approximate the data. Note that this does not exclude tuning-parameters.

3.3.1 Kernel Regression: Nadaraya-Watson

As described in section **??**, we aim to estimate

$$\mathbb{E}[Y \mid T = t] = \int_{\mathbb{R}} y f_{Y|T}(y \mid t) dy = \frac{\int_{\mathbb{R}} y f_{T,Y}(t, y) dy}{f_T(t)},$$

where $f_{Y|T}, f_{T,Y}, f_T$ denote the conditional, joint and marginal densities. This can be done with a kernel K:

$$\hat{f}_T(t) = \frac{\sum_{i=1}^{n} K\left(\frac{t-t_i}{h}\right)}{nh}, \quad \hat{f}_{T,Y}(t, y) = \frac{\sum_{i=1}^{n} K\left(\frac{t-t_i}{h}\right) K\left(\frac{y-Y_i}{h}\right)}{nh^2},$$

where h, the bandwidth, symbolizes the window size of observations to consider. By using the above function in equation (**??**), we arrive at the Nadaraya-Watson kernel estimator (NW):

$$\hat{m}(t) = \frac{\sum_{i=1}^{n} K\left((t - t_i)/h\right) Y_i}{\sum_{i=1}^{n} K\left((t - t_i)/h\right)}$$

Common choices for the kernel are the normal function or a uniform function (also called 'box' function).

[1] We used the python function `scipy.optimize.curve_fit`.

Choose Bandwidth

Note, that we still need to choose the bandwidth of the function. This can be done with the help of LOOCV while optimizing the RMSE. For non-equidistant data we refer to **?** where a local adaptive bandwidth selection is presented.

Advantages	Disadvantages
— Flexible due to different possible kernels.	— If the $t \mapsto K(t)$ is not continuous, $\hat{m}$ isn't either.
— Can be assigned degrees of freedom (trace of the hat-matrix).	— Choice of bandwidth, especially if t_i are not equidistant.
— Estimation of the noise variance $\hat{\sigma}_\varepsilon^2$.[2]	— Biased estimator. Underestimation (overestimation) for peaks (valleys).

3.3.2 Universal Kriging

Universal Kriging (UK) as described in **?** was developed in geostatistics to deal with autocorrelation of the response variable at locations that are spatially close. By applying the notion that two spectral indices that are timewise close should also take similar values, we justify the application of UK. In the end, we would like to fit a smooth Gaussian process to the data.

A Gaussian Process $\{S(t) : t \in \mathbb{R}\}$ is a stochastic process if $(S(t_1), \ldots, S(t_k))$ has a multivariate Gaussian distribution for every collection of times $t_1, \ldots, t_k$. S can be fully characterized by the mean $\mu(t) := E[S(t)]$ and its covariance function $\gamma(t, t') := \mathrm{Cov}(S(t), S(t'))$. Furthermore, we will assume the Gaussian process to be stationary. That is for $\mu(t)$ to be constant in t and $\gamma(t, t')$ to depend only on $h = t - t'$. Thus, we will write in the following only $\gamma(h)$.[3] Now, we need to make some assumptions on the covariance function. For this we introduce the Variogram of a Gaussian process as

$$V(h) := V(t, t+h) := \frac{1}{2}\mathrm{Var}(S(t) - S(t+h)) = \gamma(0) + \gamma(t)$$

and define γ via the above equation by choosing the Gaussian Variogram defined by

$$V(h) = p \cdot \left(1 - e^{-\frac{h^2}{\left(\frac{4}{7}r\right)^2}}\right) + n.$$

Here h denotes the distance, n is the nugget, r is the range and p is the partial sill. The influence of the parameters is visualized in figure **??**.[4]

Finally, we consider a one-dimensional Gaussian process G_γ with Variogram γ and tune the Variogram parameters using maximum likelihood[5]. Let z be a vector with the new values to extrapolate, then we can determine the values $m(z) = \mathbb{E}[G_\gamma(z)|(t, y)]$ using Bayes rule[6]. For an example fit, we refer to figure **??**.

[3] Note that the process is also isotropic (i.e., $\gamma(h) = \gamma(\|h\|)$ since we are in a one-dimensional setting and the covariance is symmetric.

[4] Strictly speaking, we use a scaled version of the Variogram. Thus, only the ratio of p/n matters.

[5] As illustrated in figure **??** maximum likelihood estimation can lead to overfitting. Thus, we will in practice sample several such optimized parameters and use their median in the end.

[6] Bayes rule generally claims that for two random variables A and B we have that $P(A|B) = P(B|A)/P(B)$.

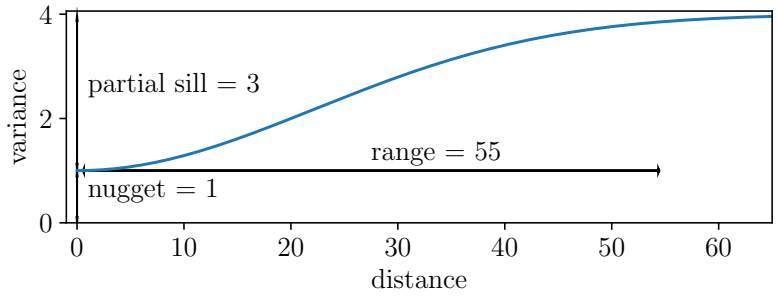


Figure 3.2: Gaussian Variogram with nugget=1, partial sill=3, range=55

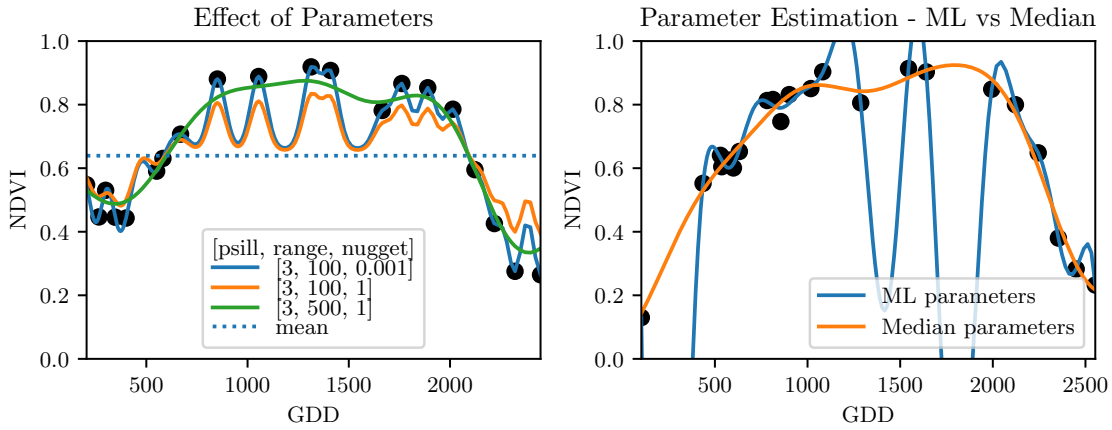


Figure 3.3: On the left, we see how the interpolation changes if we increase the nugget and the range parameter. On the right, we compare two UK interpolations, where one takes parameters by numerically maximizing the (which results in a very small nugget) and the other takes the median of many such numerical optimizations.

Violated Assumption

Since we observe a clear pattern of a growth period in spring and harvest in the end of summer, we conclude that our stationarity assumption with the constant mean is structurally violated. This is also the reason why we observe a tendency to the mean, as indicated in figure **??**, regardless of range parameter.

Advantages	Disadvantages
— It is a well-studied method.	— Violated stationarity assumption (constant mean and constant variance).
— Informative variogram parameters.	— Tendency to the mean (especially within data gaps).
— Flexible covariance structure.	— Pure maximum likelihood can result in overfitting.

3.3.3 Savitzky-Golay Filter

The Savitzky-Golay Filter (SG), introduced in **?** is a technique in signal processing and can be used as a low-pass filter (**?**). Hence, it can also be used for smoothing by filtering

high frequency noise while keeping the low frequency signal.

First, we choose a window size m. Then, for each point, $j \in \{m, m+1, \ldots, n-m\}$ we fit a polynomial of degree k by:

$$\hat{y}_j = \min_{p \in P_k} \sum_{i=-m}^{m} (p(t_{j+i}) - y_{i+j})^2,$$

where P_k denotes the Polynomials of degree k over $\mathbb{R}$. For equidistant points this can efficiently be calculated by

$$\hat{y}_j = \sum_{i=-m}^{m} c_i y_{j+i},$$

where the c_i are only dependent on the m and k and are tabulated in the original paper.

? developed a 'robust' IM variant of the SG. The method is based on the assumption that due to atmospheric effects the observed NDVI tends to be underestimated and that it cannot increase too quickly. The latter is argued by the biological impossibility of such fast vegetation changes. Their proposed algorithm is:

i.) Remove non-SCL45 points.

ii.) Remove points that would indicate an increase greater than 0.4 within 20 days.

iii.) Linearly interpolate to obtain an equidistant TS X^0.

iv.) Apply the SG to obtain a new TS X^1.

v.) Update X^1 by applying again a SG. Repeat this until $w^T|X^1 - X^0|$ stops decreasing, where w is a weight vector with $w_i = \min\left(1, 1 - \frac{X_i^1 - X_i^0}{\max_i \|X_i^1 - X_i^0\|}\right)$. This reduces the penalty introduced by outliers[7] and by repeating this step we approach the 'upper NDVI envelope'.

Extension: Spatial-Temporal SG

One notable adaptation of the SG is the presented by **?**. The key difference is the additional assumption of the cloud cover being discontinuous and that we can improve by looking at adjacent pixels[8]. Because we are working with rather high-resolution satellite data, and we need the variance in the predictors, we will waive this extension.

Advantages	Disadvantages
— Popular technique in signal processing.	— No natural way of how to estimate points that are not in the data.
— Efficient calculation for equidistant points.	— Not generalizable to other spectral indices.
— The upper envelope matches intuition for the NDVI. Therefore, it is robust against outliers with small values.	— Linear interpolation to account for missing data might not be appropriate.
	— No continious interpolation between two measurements.

[7] Here we call a point i an outlier if $X_i^0 < X_i^1$.

[8] Here, we say that a pixel is adjacent if it is the same pixel but from a different year (keeping the same day of the year) or (if not enough of such temporal-adjacent pixel are found) it is spatially adjacent.

3.3.4 Locally Weighted Regression

The Locally Weighted Regression (LOESS) introduced by **?** can be understood as a generalization of the SG (cf. section **??**).

Given a proportion $\alpha \in (0, 1]$, we estimate each y_i separately by fitting a polynomial of order d by weighted least squares. The weights are (usually) defined by

$$w_i(t_j) = \begin{cases} \left(1 - \left(\frac{|t_j - t_i|}{h_i}\right)^3\right)^3, & \text{for } |t_j - t_i| < h_i \\ 0, & \text{for } |t_j - t_i| \geqslant h_i \end{cases},$$

where h_i is the minimal distance such that $\lceil \alpha n \rceil$ observations are in the ball $B_{h_i}(t_i)$.[9] So for each y_i we only consider a proportion α of the observations.

Differences between the Robust LOESS and the SG

The LOESS smoother takes a fraction of points instead of a fixed number, and therefore automatically adapts to the size of the data we wish to interpolate. However, we run into the danger of considering too little observations, since the estimation breaks down if $\lceil \alpha n \rceil < d + 1$.[??] Furthermore, LOESS gives less weight to points further away. This yields a 'smoother' estimate, since when we slide the window (e.g., for estimating the next value) an influential point at the border does not suddenly get zero weight from being weighted equally before. Finally, the LOESS can also be used for non-equidistant data and allows for arbitrary interpolation.

Advantages	Disadvantages
— Flexible generalization of SG. — Arbitrary interpolation possible. — Intuitive parameters.	— The nature of local regression might lead to surprising estimates (no smoothness guarantees for the second derivative).

3.3.5 B-Splines

B-Splines (BS) as discussed in **?** are piecewise cubic polynomials defined by

$$S(t) = \sum_{j=0}^{n-1} c_j B_{j,k;t}(t),$$

where B are basis functions and recursively defined by:

$$B_{i,0}(z) = 1, \text{ if } t_i \leq z < t_{i+1}, \text{ otherwise } 0$$
$$B_{i,k}(z) = \frac{z - t_i}{t_{i+k} - t_i} B_{i,k-1}(z) + \frac{t_{i+k+1} - z}{t_{i+k+1} - t_{i+1}} B_{i+1,k-1}(z).$$

Assuming that all t_i are distinct, this yields an interpolation that fits the data perfectly. To reduce the amount of overfitting and increase the smoothness, we relax the constraint

[9] If too many weights are set to zero, we might end up considering not enough observations and thus get a singular design-matrix (for the least squares estimation). Therefore, we substitute h_i with $1.01h_i$, so that the observation on the boundary of $B_{h_i}(t_i)$ does not get completely ignored. But we also have to assure that α is big enough.

that we must perfectly interpolate. Thus, we use the minimum number of basis functions[10] such that:

$$\sum_{i=1}^{n}(w_i(y_i - \hat{y}_i))^2 \leq s$$

Advantages	Disadvantages
— Can be assigned degrees of freedom. — Extendable to 'smooth' version. — Also performs well if points are not equidistant.	— Smoothing process does not translate well to an interpretation (unlike SS). — Choice of smoothing parameter s.

3.3.6 Smoothing Splines

To define the Smoothing Splines (SS) as in **?**, let $\mathcal{F}$ be the Sobolev space (the space of functions of which the second derivative is integrable). Then the SS are the unique[11] minimizer of

$$\hat{m} := \underset{f \in \mathcal{F}}{\arg\min} \sum_{i=1}^{n} w_i \left(y_i - f\left(t_i\right)\right)^2 + \lambda \int f''(t)^2 dt$$

The objective function ensures that we decrease the curvature while keeping the RMSE low. SS can be represented as cubic splines i.e., piecewise cubic polynomial functions.

Whittaker — Discrete Version with Higher Order Derivatives

The Whittaker smoother introduced in **?** is closely reminiscent of the SS and is also used for the NDVI TS (**?**). Similar to SS, we minimize the following expression over $z \in \mathbb{R}^n$:

$$(y - z)^T W (y - z) + \lambda z^T D^T D \mathbf{z},$$

where W is a diagonal weight-matrix, λ our parameter and D a matrix that serves the purpose of approximating a differentiation of k-th order. In essence, this minimization function is the same as equation **??**. The only differences are, that we substitute the integral by a sum and that we are more flexible with the order of the derivatives we are using. The main drawback is that we do not get a smooth function that interpolates, and that the sum behaves worse than the integral for non-equidistant data points. Thus, we will not consider the Whittaker further but consider the more general SS.

[10] So we do not require one basis function for each neighboring pair of knots. SciPy uses FITPACK and DFITPACK, the documentation suggests that smoothness is achieved by reducing the number of knots used.

[11] Strictly speaking, it is only unique for $\lambda > 0$.

Advantages	Disadvantages
— Can be assigned degrees of freedom (trace of the hat-matrix). — Efficient estimation (closed form solution). — Intuitive penalty (we don't want the function to be too 'wobbly' — change slopes). — Also performs well if points are not equidistant. — Fixes the Runge's phenomenon (fluctuation of high degree polynomial interpolation). — Bounded within the data range if λ is chosen a priori.	— The tuning parameter λ must be chosen. This can be done via cross validation and optimizing a score function (e.g., the RMSE).

3.4 Tuning Parameter Estimation

Many of the IMs introduced in section **??** and **??** include a free parameter. To determine this parameter for a specific IM, we will estimate the absolute residuals using OOB estimation and then optimize the parameter using a score function. We clarify the procedure step by step:

i.) Construct a set Λ of candidate parameters that generously covers the parameter space.

ii.) Consider the vector $z := (y_1^{(1)}, \ldots, y_{n_1}^{(1)}, y_1^{(2)}, \ldots, y_{n_2}^{(2)}, \ldots, y_1^{(m)}, \ldots, y_{n_m}^{(m)},)$ of all NDVI TS for all m pixels.

iii.) Consider its LOOCV estimator $\hat{z}_\lambda$ using the specific IM and the parameter λ for each pixel individually.

iv.) Determine $\lambda_{optimal} = \arg\min_{\lambda \in \lambda} \mathrm{QAR}^{90}(\hat{z}_\lambda)$

We choose QAR^{90} as our optimization function because we want to allow 10% of outliers (contaminated points) but also aim for an accurate fit in 90% of the cases.

Figure **??** exemplifies the effect of the chosen quantile in the optimization function ($\mathrm{QAR}^{\mathrm{x}}$). To summarize, we may say that the higher the quantile, the stronger the smoothing.

3.5 Robustification

Now we discuss a general approach of how to make an interpolation more robust against outliers. The main idea is to give less weight to observations that have high residuals after the initial (or if we reiterate, the previous) fit.

Even though the procedure is taken from the robust version of the LOESS smoother (cf. section **??** and **?**), we can apply it to every IM that allows for prior weighting of observations.

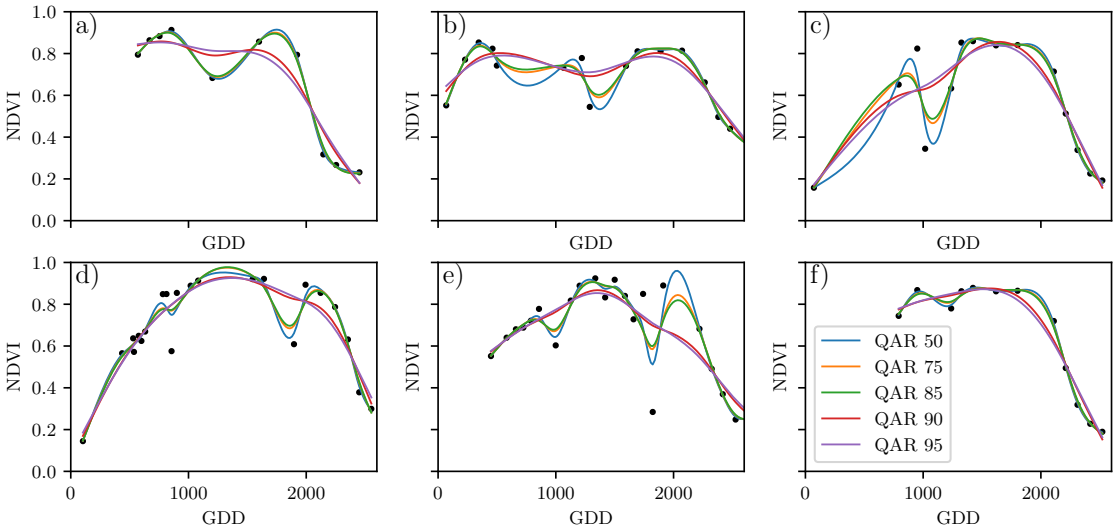


Figure 3.4: Smoothing splines fit with smoothing parameter optimized by minimizing the LOOCV QAR. Note that the larger the considered quantile is, the more the curvature of the resulting curve decreases.

After an initial fit, we calculate the residuals $r_i := y_i - \hat{y}_i$ and obtain $\tilde{r}_i$ by scaling with the median of the absolute residuals:

$$\tilde{r}_i := \frac{r_i}{6 \operatorname{med}(|r_1|, \ldots, |r_n|)}$$

Next, we compute new weights by

$$w_i^{\text{new}} := w_i^{\text{old}} \begin{cases} \left(1 - \tilde{r}_i^2\right)^2, & \text{if } |\tilde{r}_i| < 1 \\ 0, & \text{else} \end{cases};$$

Using the new weights, we can re-interpolate. This reweighting can be iterated for several steps or till the change of the values is smaller than some tolerance.

Note that this procedure is indeed robust since we use the median for the normalization which has a breakdown point[12] of 50%.[13]

3.5.1 Our Adjustment:

During the iterations or when supplying prior weights, low-weighted observations can corrupt our estimation of scale (the median of absolute residuals). Thus, we introduce the weighted median as

$$\operatorname{med}_{\text{weighted}}(r, w) := \arg\min_{\lambda \in \mathbb{R}} \sum_{i=1}^{n} |r_i w_i - \lambda|$$

for $r, w \in \mathbb{R}^n$.

[12] Intuitively, the breakdown point denotes the fraction of observations a 'vicious' player can replace without breaking the estimator. For example, the median has a breakdown point of 50%.

[13] The breakdown point relates only to outliers in the y values. Note that we do not require the IMs to be robust, since the residual for an outlier will still be larger than for non-outliers and thus will be down weighted more and more in each iteration (because for the next iteration the residual of the outlier will be even larger, since we gave less weight to it).

3.5.2 Examples and Conclusions

Examples of the first four iterative fits using SS and LOESS are shown in figure **??** and **??** for six pixels. For the analogous figures of BS and DL we refere to the appendix **??** and **??**. Indeed, we observe how the interpolated TS is less affected by outliers after each iteration. We notice the biggest difference in the first iteration. Furthermore, in the plot at the bottom left we see how the interpolation 'escapes' from the right endpoint with each successive iteration, even though our intuition does not necessarily identify this point as an outlier. Therefore, in the following, we will always stop after one iteration.

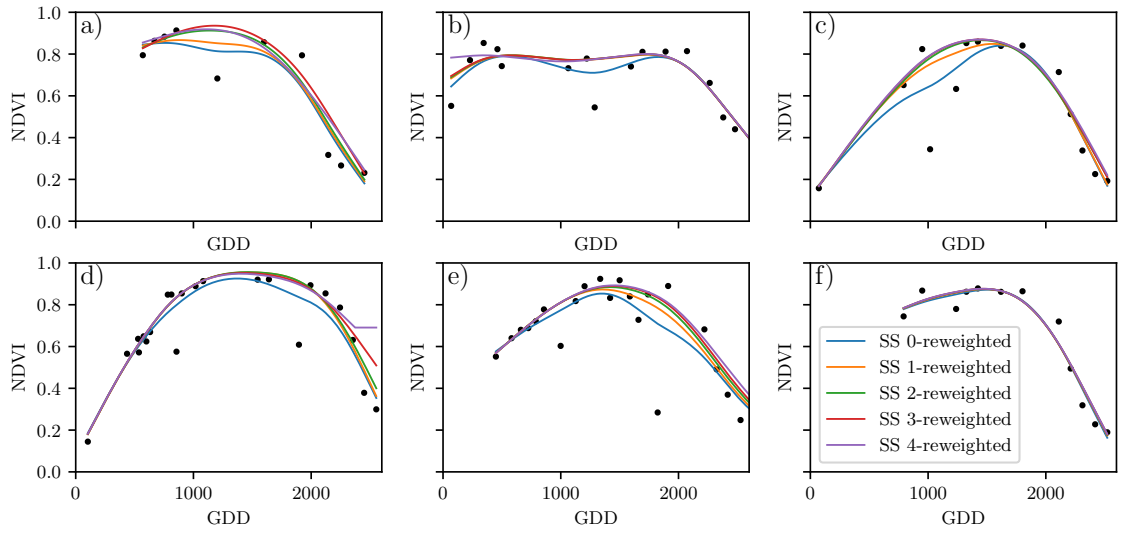


Figure 3.5: Smoothing splines fitted to different (SCL45) NDVI TS. Iterations of a robustifying refit (as indicated in section **??**) are also displayed.

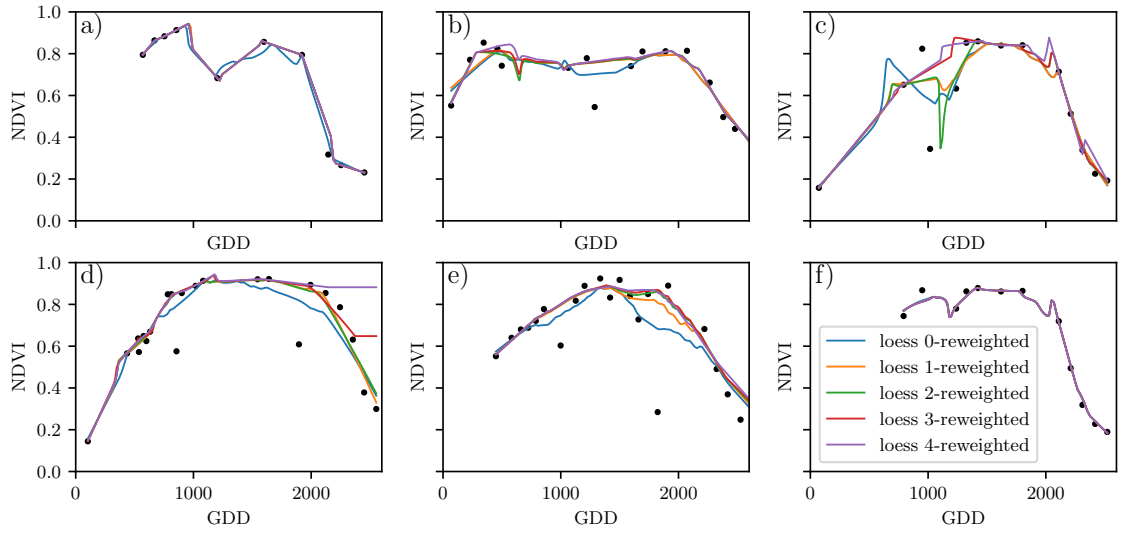


Figure 3.6: The LOESS smoother fitted to different (SCL45) NDVI TS. Iterations of a robustifying refit (as indicated in section **??**) are also displayed.

3.5.3 Upper Envelope Approach

If we artificially increase the negative residuals in **??** (e.g., by multiplying with 2), the corresponding points will get less weight in the next iteration. This allows us to create an interpolation that resembles an upper envelope. Intuitively, this upper envelope can be thought of as a sheet that is laid on top of the points.

This approach is based on the premise that we tend to underestimate the NDVI (**?**). Since we want to develop a general method that is in principle not related to the NDVI, we will not pursue this approach further.

3.6 Performance Assessment

Next, we will benchmark the in section **??** preselected IMs with and without robustification. For this, we will use the same technique as we did for the parameter determination in section **??**. On B_λ we apply the score functions from section **??**.

The results are presented in section **??** and are discussed in section **??**. The double logistic turns out to be the best convincing parametric method, and from the non-parametric methods we choose the SS.

Table 3.1: Summary of the studied interpolation methods containing important assumptions, advantages and disadvantages and whether the method supports weighted observations (w) and if the resulting interpolation is bounded w.r.t. a fixed interval (b).

	Assumptions	**Advantages**	**Disadvantages**	**w**	**b**
Double-Logistic	– Bell shape of curve – NDVI has a minimal value	– Good for evergreen plants (if snow masks NDVI) – Handles data gaps well	– Parameter estimation can be challanging, to solve this the parameter space can be bounded	Yes	(Yes)
Fourier Series	– NDVI can be approximated by a 2^{nd} order Fourier series. – Prominent approach	– Incorporates periodical growth-cycles	– Curve easily exceeds the bounds of the NDVI within data gaps – Parameter estimation can be challanging, to solve this the parameter space can be bounded	Yes	No
Nadaraya-Watson (Kernel Smoothing)	– Close points are related to each other via a kernel function	– Simple – Computationally very fast	– Biased, especially at 'peaks' and 'valleys' – Bandwidth: fails if there are big data-gaps	Yes	Yes
Universal Kriging	– Function is a realization of a stationary Gaussian process	– Informative parameters – Flexible	– Assumption not met for NDVI TS – Regression to the mean, especially within data gaps	Yes	(Yes)
SG	– High frequencies are noise (Low-Pass-Filter) – Equidistant points – Local polynomials	– Computationally very fast	– Cannot deal natively with non-equidistant data	No	(Yes)
SG + NDVI	– Upper envelope – Vegetation cannot grow faster than some slope	– Biological knowledge	– Bad 'upper envelope' since weights are not used for the estimation itself	(No)	(Yes)
LOESS	– Local polynomial with points closer to the estimated point are more important	– Flexible – Generalization of SG – Intuitive weighting function	– Computationally expensive	Yes	(Yes)
B-Splines (Smoothed)	– Function can be approximated by a linear combination of B-splines basis functions	– General assumption – Flexible shape	– Unbounded – Non-intuitive smoothing process	Yes	No
Smoothing splines	– 2^{nd} derivative of function is integrable	– Intuitive meaning of penalty – General assumptions – Flexible shape	– Choice of smoothing parameter	Yes	(Yes)

Chapter 4

NDVI Correction

Let's remind ourselves that the data from the S2 satellites is distributed with a SCL, and we therefore have some evidence about what is observed at each pixel for each sampled time (cf. table **??**). So far, we have only considered points, labeled as cloud- and shadow-free (SCL45). However, we remind ourselves of the satellite images in figure **??**, where we had cloudy images despite the 'vegetation' label and see vegetation in figure **??** even though we are supposed to see 'cirrus clouds'.

In this chapter, we will try to improve our NDVI interpolation by not relying only on the observed NDVI, but by training our own model to correct the NDVI using all S2 bands. For this, we introduce several statistical modelling approaches and discuss the strengths and weaknesses for each of them. After correcting the observed NDVI, we will assess the uncertainties of our corrections and translate them into weights. These will be used for the subsequent interpolation. This step-by-step procedure is illustrated in figure **??** in the appendix. Finally, we will evaluate which combinations of IMs and correction models perform the best.

4.1 Considering other SCL Classes

In figure **??** we plot the observed NDVI and notice that some blue points which correspond to the SCL-class 10 (thin cirrus clouds) follow the interpolated line closely. Hence, they might be useful in improving an interpolation fit.

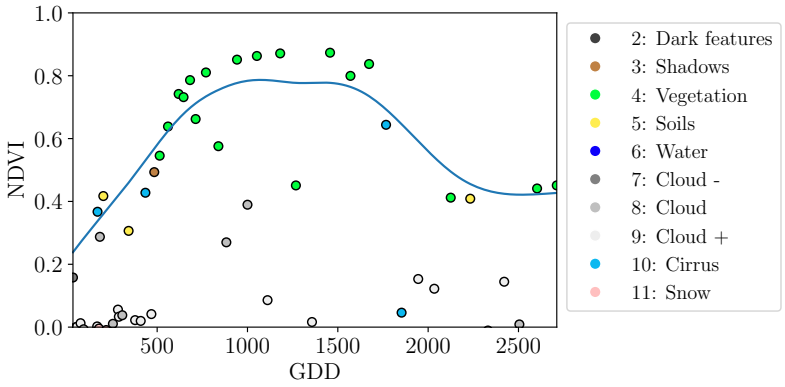


Figure 4.1: A smoothing splines fit considering green and yellow points (SCL45).

To get an impression of whether there is some useful information contained in non-SCL45 observations, we compare the observed NDVI with the true NDVI. But since we do not have any ground truth data, we will make the following assumption:

Assumption 4.1.0.1. The 'true' NDVI value at time t can be successfully estimated by a robustified LOOCV interpolation using high-quality observations. That is, the interpolated value (using a robustified IM from chapter **??**) considering the points $P^{SCL45} \setminus P_t$. In the following, we will call this estimate the 'true'-NDVI.

We would like to get an idea if there is any information that can be recovered from non-SCL45 observations. For that, we will check for the other SCL-classes if there is a relation between the 'true' NDVI (derived with robustified SS) and the observed NDVI. Thus, we pair each 'true' NDVI with its observed one, collect all pairs, and create a scatter plot for each SCL-class in fig **??**. As expected, the 'true' and the observed NDVI seem to be highly correlated for SCL45. But we can also detect some patterns of correlation in the SCL-classes 2, 3, 7, 8 and 10.

It might be tempting to just include some of the mentioned SCL classes for the interpolation. But on the one hand, the choice would not be objective and on the other hand, the correlation seems to be weaker than for SCL45. Therefore, in the following section, we will correct the observed NDVI and estimate the uncertainty of each correction.

4.2 Correction Models

For training an NDVI correction model, we require ground-truth data which we will aim to model using informative covariates. Since ground-truth NDVI data is not available, we will again use the assumption **??** and use the 'true' NDVI instead. There is no canonical answer to the question of which covariates we should use. It is a tradeoff between simplicity, generalizability, and performance (with the danger of overfitting). Our goal with the NDVI correction is to develop a product that is simple to use and to understand. Therefore, in the subsequent, we will only take the spectral data of the satellite (i.e., all the bands) and the observed NDVI derived from it as covariates. We organize the chosen covariates in the design matrix X[1], where each row corresponds to a P_t (i.e., a pixel at a time t) and each column to one covariate.

In the following, we will introduce different approaches to model the relationship between the response $y := \text{NDVI}_{\text{'true'}} \in \mathbb{R}^n$ and the design matrix $X \in \mathbb{R}^{n \times p}$. First, we will study the basic OLS. Second, we look at the LASSO, a penalized adaptation of the OLS which is known to successfully deal with highly correlated covariates. Afterwards, GAMs are introduced, which model the response similar to OLS but allow for non-linear relations. Last, we discuss RF and MARS, which are both flexible modelling approaches.

Note that in order to reduce computation time, only 10% of the data has been used to fit the subsequent models, which are still more than 120'000 observations.

4.2.1 Ordinary Least Squares

The Ordinary Least Squares estimator (OLS) is a linear model that aims to minimize the sum of the squared residuals. We assume a linear relationship between y and X and allow

[1] Strictly speaking, we include also the intercept and introduce one dummy variable for each SCL-class.

for Gaussian noise. That is:

$$y = X\beta + \epsilon \quad \text{where } \epsilon \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$$

Assuming that $(X^T X)$ is regular, we can estimate the regression coefficients β by

$$\hat{\beta} = (X^T X)^{-1} X^T y = \underset{\beta in \mathbb{R}^p}{\arg\min} \|y - X\beta\|_2^2$$

We will train two models, one using all covariates discussed above and one using only the SCL-classes and the observed NDVI.

Advantages	Disadvantages
— Simple method with good interpretability of coefficients. — Computationally cheap.	— Catches only linear relationships. — No integrated variable selection.[2]

4.2.2 Least Absolute Shrinkage and Selection Operator

The Least Absolute Shrinkage and Selection Operator (LASSO) can be similarly expressed than the OLS but adds a penalty to the minimization problem:

$$\hat{\beta}_\lambda = \underset{\beta \in \mathbb{R}^p}{\arg\min} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 = \underset{\beta \in \mathbb{R}^p \text{ and } \|\beta\|_1 < \lambda}{\arg\min} \|y - X\beta\|_2^2.$$[3]

Even though we do not have a closed form solution for equation (**??**) we can solve it easily via optimization, since the function $\beta \in \{\beta \in \mathbb{R}^p | \|\beta\|_1 < \lambda\}\| \mapsto \|y - X\beta\|_2^2$ is continuous and convex.

? shows that the LASSO solution tends to be sparse. That is $\beta_i = 0$ for most $i = 1, \ldots, p$. The larger λ, the more $\beta_i = 0$ and hence the simpler the resulting model.

In order to know which λ to choose, we try a huge range of possible values. For each β_λ, we calculate the cross-validated $RMSE_\lambda$ [4] (and its standard deviation σ_λ using the k folds) and define the λ with the smallest corresponding $RMSE_\lambda$ as λ_{min}. From here we choose the largest λ for which the $RMSE_\lambda$ is smaller than $RMSE_{\lambda_{min}} + \sigma_\lambda$. This yields a simpler model while keeping the $RMSE$ reasonable model.

We will apply the LASSO using the selected covariates in section **??** and their second degree of interactions.[5]

Advantages	Disadvantages
— Usually yields a sparse solution. This tends to give better generalizability (prediction performance on unseen data). — Successfully deals with correlation in covariates. — Interpretable results.	— Estimate is biased. — Computationally expensive.

[3] The last two terms are equivalent by lagrangian optimization.

[4] The cross validated Root Mean Square Error is the mean of the RMSE's obtained for each fold using the model trained on the remaining folds.

[5] This is if our covariates are $\{1, a, b\}$, then we will now use $\{1, a, b, ab, a^2, b^2\}$.

4.2.3 General Additive Model

General Additive Models (GAM) as described in **?** are a special case of Projection Pursuit Regression, where only the p directions parallel to the coordinate axes are considered. The result is different to a linear model since the coordinate functions are not restricted to be linear but are assumed to be non-parametric functions. The model can be written as:

$$g_{add}(x) = \mu + \sum_{i=1}^{p} g_j(x_j).^{6}$$

To estimate the non-parametric functions, we can use SS (cf. section **??**). For this let $\mathcal{S}_j$ be the function that takes some $z \in \mathbb{R}^n$ and returns the SS fitted to $(X_{:,j}, z)$ where the smoothing parameter is optimized by LOOCV[7]. Since we cannot fit all g_j simultaneously, we will use a strategy named Backfitting. We basically cycle through the indices $1, \dots p$ and refit $\hat{g}_j$ each time. The following illustrates the procedure:

$$\begin{aligned}
&1) \quad \hat{g}_1 &&= \mathcal{S}_1(y - \mu) \\
&2) \quad \hat{g}_j &&= \mathcal{S}_j(y - \mu - \hat{g}_1(X_{[:,1]}) - \cdots - \hat{g}_{j-1}(X_{[:,j-1]})) \quad \text{for } j = 2, \dots, p \\
&3) \quad \hat{g}_1 &&= \mathcal{S}_1(y - \mu - \hat{g}_2(X_{[:,2]}) - \cdots - \hat{g}_p(X_{[:,p]})) \\
&4) \quad \hat{g}_j &&= \mathcal{S}_j(y - \mu - \sum_{k \neq j} \hat{g}_k(X_{[:,k]})) \quad \text{for } j = 2, \dots, p \\
&\quad \vdots
\end{aligned}$$

We repeat step 3) and 4) until the change falls below some tolerance.

Advantages	Disadvantages
— Captures non-linearity.	— No automatic variable selection.
— Good interpretability.	— Computationally expensive.

4.2.4 Random Forest

To define a Random Forests (RF) introduced by **?** we will first define what a Tree is. A (decision) Tree is a graph (V, E) without circles, a distinct root node, every node has at most two children and every leaf has a value assigned to it. At each node there is a boolean condition testing if one variable is greater than some value and a pointer to one child depending on the boolean value. To evaluate a tree we start at the root node, test the boolean expression and go to the node indicated by the resulting pointer. This we repeat until we end up at a leaf-node, where we return the value assigned to it.

To build such a Tree, we will recursively partition the covariate space using greedy splits[8] decreasing the RMSE[9] each time. If the set we want to split contains less than a certain amount of training points, we stop.

[6] Where g_j is a real-valued function. For identifiability we also demand $\mathbb{E}[g_j(X_{:,j})] = 0$ for $j = 1, \dots, p$.

[7] For efficiency a proxy of the LOOCV is used called generalized cross validation.

[8] For computational reasons, we will only use splits along one covariate. So we 'cut' our covariate space into rectangles.

[9] To calculate the RMSE, we need a prediction. Let P be the current partition, then the predicted value for some $x \in A \in P$ is the mean of the responses of all the points in A (included in the training data).

To build a Random Forest we will bootstrap-aggregate[10] many such Trees[11]. The prediction of the Random Forest for a new point x is then the mean of the predictions from all the Trees.

Advantages	Disadvantages
— Captures non-linear relationships.	— The resulting (prediction) function is not continuous, but locally constant.
— Captures all interactions and performs automatic variable selection.	— Computationally expensive.
— Can deal with missing data.	— No interpretability.

4.2.5 Multivariate Adaptive Regression Splines

A Multivariate Adaptive Regression Splines (MARS) model as introduced in **?** can be described by

$$g(x) = \sum_{m=0}^{M} \beta_m h_m(x),$$

where the h_m are simple functions (explained later) and the β_m are estimated via Least Squares.

In the building procedure of a MARS model, we first select many of those simple functions and later drop some of them to avoid overfitting. For the construction of those simple functions, define $\mathcal{B}$ be the set of pairs of 'hockey stick functions'

$$\mathcal{B} := \left\{(b_1, b_2) \mid (b_1(x), b_2(x)) = \left((x_j - d)_+, (d - x_j)_+\right), d = X_{1,j}, \ldots, X_{n,j},\ j = 1, \ldots, p\right\}$$

and the set $\mathcal{M} = \{1\}$ of all functions currently in the model. Now, consider $\mathcal{C}$ the set of candidate functions-pairs

$$\mathcal{C} := \{(h(\cdot)b_1(\cdot), h(\cdot)b_2(\cdot)) \mid h \in \mathcal{M},\ (b_1, b_2) \in \mathcal{B}\}$$

and select the pair (which when added to $\mathcal{M}$ and the coefficients refitted) reduces the RMSE the most. Add the selected pair to $\mathcal{M}$ and repeat until the RMSE reduction becomes insignificant.

Finally, to avoid overfitting, we prune the set $\mathcal{M}$ by optimizing a LOOCV score.[12]

To reduce computational complexity, we follow the recommendation from **?** and restrict h in equation (**??**) to be of degree one (so it is also in a pair of $\mathcal{B}$). Consequently, $\mathcal{C}$ contains functions with a degree of at most 2.

[10] That is we will sample (with replacement) several times n observations from our original data and fit a Tree to each such sample.

[11] Building the Tree, this time we will not test every covariate at each node (for the RMSE minimization) but a node-specific subsample of the covariates. Thus, also the 'second best split' can be selected.

[12] This means that we perform an iterative procedure to reduce the number of functions in $\mathcal{M}$. For every function h in $\mathcal{M}$, we compute the model using $\mathcal{M}$
$\{h\}$. We discard the function that – when excluding from $\mathcal{M}$ – leads to the best LOOCV score.

Advantages	Disadvantages
— Catches non-linear relationships. — Interpretability via functions in $\mathcal{M}$ and their coefficients. — Allows for interactions with variable selection.	— Computationally expensive (can be reduced by restricting the degree of interactions).

4.3 Weighted Interpolation

Once we have corrected the NDVI using the models described in the previous section, we are left with the problem that not every correction is equally reliable.[13]. Hence, we are interested in a measure of how uncertain an estimate is. We achieve this analogously as we corrected the NDVI, by replacing the response $\text{NDVI}_{\text{`true'}}$ with the absolute residuals $v := |y - \hat{y}|$ and modeling their relationship with the covariates defined by X. In this way, we obtain a model for the absolute residuals v and the estimator $\hat{v}$.

In the following, we will convert our uncertainty estimate into weights that can be used for interpolation. For this, consider a pixel P, $\hat{y}^{(P)}$ its corrected NDVI values and $\hat{v}^{(P)}$ the estimated uncertainties of $\hat{y}^{(P)}$. In order to interpolate $\hat{y}^{(P)}$, we will give less weight to unreliable observations. Thus, we define the link function connecting $\hat{v}$ with weights:

$$w_\tau^{(P)} := \frac{1}{R}\frac{1}{\hat{v}_\tau^{(P)}}, \qquad \text{for } \tau = 1, \dots, n_P$$

where τ is an index over the satellite images and $R := \frac{\sum_i^{n_P} \hat{v}_i^{(P)}}{n_P}$ a normalization constant. The normalization is needed since for some IMs, inflating the sum of weights would decrease the effect of the smoothing.

4.4 Resulting Interpolation Strategies

We have developed the following procedure to obtain a new interpolation (keyword-wise):

i.) LOOCV Interpolation (+ robustify?) to get 'true' NDVI

ii.) Correction

iii.) Uncertainty estimation

iv.) Interpolation (+ robustify?)

At each step we have a choice, more precisely:

— Interpolation: SS / DL

— Robustify: Yes / No

— Correction & uncertainty estimation: RF / OLS – considering only SCL-classes / OLS – considering all selected covariates / MARS / GAM / LASSO / no correction.

As it is not feasible to try every possible combination, we make the following restrictions on which combinations we will consider:

[13] One correction is illustrated in figure **??**. In this figure, the outer points (labeled as clouds) have a large scatter.

— We use the same IM each time.

— Either we robustify both times, or we do not robustify at all.

— We use the same underlying method for correction and uncertainty estimation.

In this fashion, we obtain 28 distinct Interpolation Strategies (ISs), which we will benchmark in the next section.

4.5 Evaluation via (relative) Yield Prediction Error

In this section, we introduce the (relative) Yield Prediction Error (relative YPE) and utilize it to evaluate the 28 ISs from section **??**. The fundamental assumption is that the closer the interpolated NDVI TS is to the true one, the better it can be used to determine crop yield. Implicitly, we believe that an NDVI TS that better models yield will incorporate more true information about the underlying vegetation. Therefore, we want to determine a comparable YPE for each IS and choose it as a benchmark criterion. This is an objective measure since we have not considered crop yield in any of our previous steps. Moreover, this criterion is justified by the fact that yield estimation has been a motivation for the interpolation.

Definition 4.5.0.1. *(Relative YPE) Let $y \in \mathbb{R}^n$ be the yield, M be a model for estimating y, and $\hat{y} = M(X)$ where X describes the data*[14]*. We define the relative YPE as the relative RMSE in yield estimation. Formally expressed:*

$$YPE = \frac{\sqrt{\sum_{i=1}^{n}(y_i - \hat{y}_i)}}{\bar{y}},$$

where $\bar{y}$ denotes the sample mean. For the (non-relative) YPE do not divide by $\bar{y}$.

We would like to estimate the yield from the NDVI TS produced by all the ISs for all pixels. However, given the high dimensionality and different lengths of the interpolation (not every TS has the same start and end point), we must first map each NDVI TS into a low-dimensional vector space of covariates. For this, we will use the following statistics:

— Maximum slope
— Minimum slope
— Integral[15] over all
— Peak (i.e., maximal NDVI)
— GDD for the Peak
— Integral[??] up to the peak
— Integral[??] after peak
— Integral[??] from 0-685 GDD
— Integral[??] from 685-1075 GDD

For the choice we were inspired by (cf. table 2 in **?**). However, we deliberately omit any statistic that involves the minimum (e.g., the NDVI-range), since we regard the minimum as a very error-prone measure due to the large influence of clouds in the TS.

As a result, for each IS, a matrix is obtained in which each row corresponds to a pixel and both the yield and the covariates (computed by applying the above statistics) are contained. Using this matrix, we train a random forest for yield estimation, and compute

[14] We will use the matrixes derived in section **??**.

[15] We will only consider the integral of the function $max(0, NDVI - 0.3)$, where 0.3 is assumed to be a minimal NDVI value (cf. satellite images **??** and **??** with their NDVI in plot **??**).

the integrated OOB[16] estimates $\hat{y}$. Note that the choice of the modeling approach does not matter much, as long as it is general enough (i.e., able to approximate any function) and we use the same one for each IS. Finally, for each IS, we calculate the YPE and describe the results in section **??**.

[16] By the integrated OOB estimates, we denote the predictions for each pixel where only trees are used, where the pixel has not been used (as n_{tree}, the number of Trees, grows the fraction of trees which do not contain a certain pixel converges to $\frac{1}{e}$).

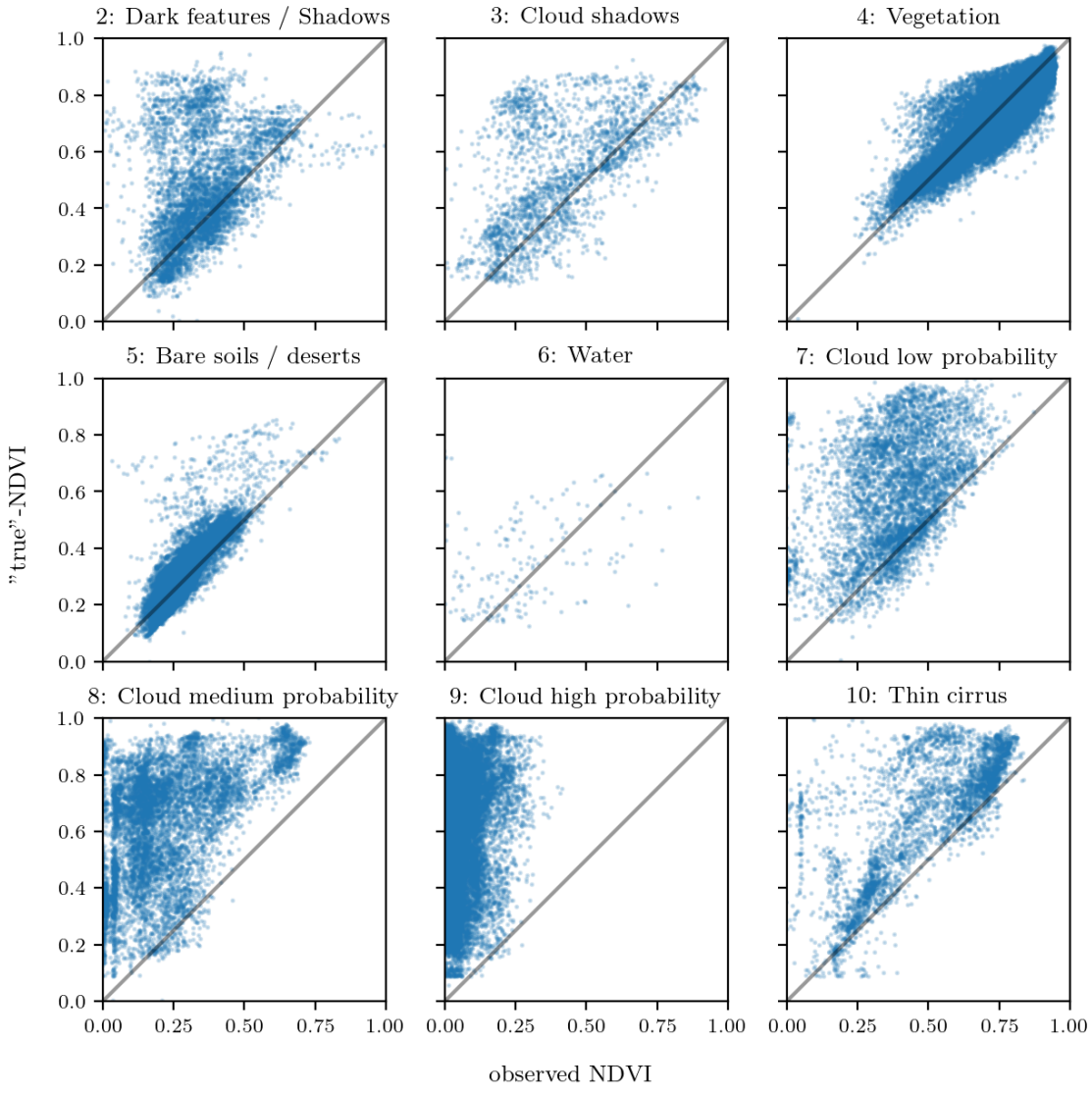


Figure 4.2: For each SCL class, we compare the true NDVI with the observed NDVI. (The true NDVI was estimated with LOOCV smoothing splines, and we used all observations of 10% of the total pixels.)

Chapter 5

Results

5.1 Goodness of Fit for Selected IMs

The benchmarks of the selected[1] IMs (on P^{SCL45}) with respect to various score functions are displayed in table **??**. The score functions summarize the absolute values of the LOOCV residuals (the smaller, the better). For each of the 5 selected IMs, we consider the basic and the robustified (see section **??**) version.

Table 5.1: Comparing the goodness of fit for selected IMs (on P^{SCL45}) measured with score functions (see section **??**) that take the LOOCV residuals as input. Colored row-wise.

	SS	LOESS	DL	BS	FS	SS^{rob}	$LOESS^{rob}$	DL^{rob}	BS^{rob}	FS^{rob}
RMSE	0.063	0.061	0.061	0.074	0.075	0.070	0.065	0.065	0.079	0.208
QAR^{50}	0.036	0.034	0.027	0.043	0.031	0.032	0.031	0.022	0.037	0.049
QAR^{75}	0.063	0.061	0.051	0.077	0.058	0.061	0.057	0.044	0.070	0.099
QAR^{85}	0.080	0.079	0.070	0.098	0.083	0.081	0.076	0.063	0.094	0.158
QAR^{90}	0.092	0.092	0.088	0.112	0.108	0.097	0.090	0.082	0.113	0.226
QAR^{95}	0.119	0.115	0.122	0.142	0.161	0.132	0.115	0.124	0.157	0.375

DL performs the best among both robustified and non-robustified with respect to most of the score functions used (all except QAR^{95}) and is in particular superior to the other parametric approach, which is FS. Especially the robust FS performs poorly. The LOESS is superior on every score function than all other non-parametric methods, but is closely followed by the SS. The BS exhibits the worst performance out of all non-parametric method tested here.

5.2 Yield Prediction Error for Tested ISs

The YPE for the in section **??** chosen ISs is given in table **??**. We note that robustification does not improve the quality of the fit (measured via the YPE) in most cases. In addition, SS tend to be better than DL (with and without robustification) in terms of YPE, especially if no correction is made. The IS that leads to the lowest YPE is the OLS^{SCL} with SS. Given that the OLS^{SCL} models have very good interpretability, we also present the regression

[1] For the discussion which IMs have been selected cf. section **??**.

Table 5.2: Relative YPE for various ISs. For the non-relative YPE and the coefficient of determination (R^2) cf. table **??** and **??**.

	RF	OLS^{SCL}	OLS^{all}	MARS	GAM	LASSO	no corrections
SS	0.155	0.140	0.143	0.142	0.142	0.142	0.149
SS^{rob}	0.155	0.143	0.147	0.149	0.146	0.145	0.148
DL	0.156	0.151	0.152	0.152	0.149	0.149	0.158
DL^{rob}	0.157	0.153	0.152	0.145	0.148	0.150	0.157

equations below. The corrected NDVI is calculated using

$$\begin{aligned}\text{NDVI}_{\text{corr}} =& 0.711\,\text{NDVI}_{\text{observed}} + 0.215\,\mathbb{1}_{SCL=2} + 0.237\,\mathbb{1}_{SCL=3} + 0.210\,\mathbb{1}_{SCL=4} \\ &+ 0.116\,\mathbb{1}_{SCL=5} + 0.162\,\mathbb{1}_{SCL=6} + 0.327\,\mathbb{1}_{SCL=7} + 0.474\,\mathbb{1}_{SCL=8} \\ &+ 0.575\,\mathbb{1}_{SCL=9} + 0.306\,\mathbb{1}_{SCL=10} + 0.512\,\mathbb{1}_{SCL=11}\end{aligned}$$

where $\mathbb{1}_{SCL=2}$ is equal to one if the current observation corresponds to SCL class 2 and zero otherwise[2]. Whereas, we obtain the estimated absolute residuals by:

$$\begin{aligned}\widehat{\text{abs}}\,(\text{NDVI}_{\text{'true'}} - \text{NDVI}_{\text{corr}}) =& -0.133\,\text{NDVI}_{\text{observed}} + 0.186\,\mathbb{1}_{SCL=2} + 0.185\,\mathbb{1}_{SCL=3} \\ &+ 0.146\,\mathbb{1}_{SCL=4} + 0.089\,\mathbb{1}_{SCL=5} + 0.167\,\mathbb{1}_{SCL=6} \\ &+ 0.203\,\mathbb{1}_{SCL=7} + 0.181\,\mathbb{1}_{SCL=8} + 0.173\,\mathbb{1}_{SCL=9} \\ &+ 0.180\,\mathbb{1}_{SCL=10} + 0.172\,\mathbb{1}_{SCL=11}\end{aligned}$$

Thus, if we observe a pixel with SCL class 4 ('vegetation') but a NDVI of only 0.4, the corrected NDVI would be $0.711 \cdot 0.4 + 0.21 = 0.494$ with an estimated absolute residual of $-0.133 \cdot 0.4 + 0.146 = 0.93$. In equation **??** we notice the strongest upwards correction for SCL classes 8, 9 and 11 ('medium probability clouds', 'high probability clouds' and 'thin cirrus clouds'). The estimated absolute residuals, however, are the smallest for SCL classes 4 and 5 ('vegetation' and 'bare soil'). Furthermore, the higher the observed NDVI the lower are the estimated absolute residuals.

For the `R`-output of the `summary` function of the two models, we refer to the appendix **??**.

[2] $\mathbb{1}$ is also called an indicator function or characteristic function in mathematics.

Chapter 6

Discussion

In the first part of the discussion, we examine IMs for compatibility with data gaps, argue choices for selected IMs and discuss the choice of the score function. In the second part, we identify the best IS and discuss issues that have arisen in the context of the NDVI correction.

6.1 Interpolation

6.1.1 Data Gaps in Time Series

NW estimates the value for the time t by relating to the points near t. To determine what 'near' means, a bandwidth h is used (cf. equation **??**). This approach becomes problematic as soon as the data gaps become larger than h, since no points are left that are close to t. Using a locally adaptive bandwidth fix (**?**), we pay with a greater variability of the estimator. According to the authors, a small sample size (as we have it for the NDVI TS) worsens this situation.

Regarding the GK, we expect that due to the stationarity assumption, the interpolation will always tend to the mean if data gaps are present (cf. figure **??**).

Since the SG requires equidistant points, data gaps will break it. **?** proposes a linear interpolation to restore missing data points. However, due to the timescale transformation to GDD in section **??**, the requirement of equidistant remains an unresolved issue.

The FS interpolation can be corrupted if there are noticeable data gaps, as can be observed in figure **??**. Additionally, the poor goodness-of-fit values for the robustified variant in table **??** illustrate the unreliability of this IM method. The values in **??** are meaningful in describing the ability to cope with data gaps, since more data points are ignored during the robustification and thus data gaps are simulated.

Similarly, for SS, LOESS, DL, and BS we compare the values in table **??** between the robustified and non-robust variant. We find that the robust variant does not differ significantly from the non-robust variant (unlike as for FS). Thus, we conclude that these methods do not have systematic failures.

Regarding the LOESS, in case of data gaps, the weights can attain non-intuitive values. The result can be a strongly fluctuating behavior, as observed in figure **??** in plot (c). There, a strange peak between the first and second observation is visible. This peak

originates from local weighting. In this case, the first data point in the plot, although adjacent to the peak, is given a low weight compared to the points to the right of the peak (for estimating the value at this peak).

In our experience, the DL handles data gaps well due to strong shape assumptions (no jumps), but it may happen that the model describes the NDVI increase as abrupt. This, however, was fixed by bounding the first derivative (cf. section **??**).

6.1.2 Preselection

Here we justify our preselection of the IMs tested in section **??**. We decided against NW because of its systematic errors at peaks and valleys and poor handling of data gaps (cf. section **??**). The UK will not be considered since the underlying stationarity assumption is not met and therefore a systematic bias is introduced. On top of that, maximum likelihood parameter estimation occasionally might lead to overfitting (cf. figure **??**). Also, we do not include the SG in the next selection, since we see it as a special case of LOESS. The remaining IMs are thus SS, LOESS, DL, BS, and FS.

6.1.3 Candidate Selection

Given that DL convinces, regarding most of the selected score functions in table **??**, we will apply this method also in chapter **??**. Moreover, we see that the robustification in most cases improved the score regarding QAR^{50}, QAR^{75}, QAR^{85} and QAR^{90}. Only for the outlier-sensitive score functions (RMSE and QAR^{95})[1] we notice significant worsening (we consider the robust FS separately in section **??**). Consequently, we will also use the robustification in section **??**. In order to not only rely on the form assumptions of the DL, we further choose a non-parametric method for further consideration. Despite the LOESS slightly dominating the SS in table **??**, we choose the SS. We justify this selection with the non-smooth behavior of the LOESS in case of data gaps (see section **??**) and the good interpretability of the SS from minimizing function **??**.

6.1.4 Score Functions

In situ data generally prefered for evaluating various IMs. Being difficult to obtain, **?** generated NDVI TS by introducing random roise to an idealized NDVI curve that has been constructed by taking the mean of multiple NDVI TS of several years. Because the distribution of the noise is known, the authors do not have to deal with systematic outliers and adequately, given the circumstances, use the non-robust RMSE. On the other hand, the generated NDVI TS no longer contain the typical challenges (underestimated NDVI and data gaps). Thus, the authors test their IMs more for general interpolation reliability, and less for challenging S2-derived NDVI TS. To adequatly test our IMs for those challenges, we employ the LOOCV. **?** evaluate various IMs based solely on the RMSE. We see the choice of the RMSE for the selection of the IMs as problematic, because it is sensitive to outliers. Yet we know that systematic outliers are present in our data (cf. figure **??**). Thus, when we consider the RMSE as a score function, we do not choose the IM that disregards outliers, but is heavily influenced by them. Likewise, **?** confirms that the RMSE, in the presence of outliers, is not sufficient. **?** proposes a robust score function, a more flexible variant of the huber loss function (**?**). For simplicity's sake, we choose the robust QAR^{x}

[1] For the RMSE one outlier is enough to take away the usefulness of the statics, in the case of QAR^{95} it is enough if 5% of the data are contaminated to break the statics.

score function (cf. definition **??**). For example, QAR^{90} can easily handle up to 10% of outliers (i.e., contaminated data points) in the data.

6.2 NDVI Correction

6.2.1 IS Selection

The evaluation of various ISs via the YPE (cf. section **??**) shows that SS are better suited than DL for yield estimation. Moreover, it seems surprising that robustification tends to worsen the results, despite reducing LOOCV residuals in most cases (cf. section **??**). We conjecture that the correction models handle outliers by themselves (by correcting or down-weighting them) and thus do not benefit from an external robustification. Indeed, for OLS^{SCL} we see in equation **??** that the smaller the observed NDVI of a point, the larger the estimated residual — yielding a lower weight. This is consistent with our experience that outliers usually underestimate the NDVI. Our conjecture is consistent with the fact that if we do not correct, robustification produces a marginal improvement.

If we use the best IS without correction (SS^{rob}) the relative RMSE of the NDVI-based yield prediction is 0.148 (cf. table **??**). Using the best IS with correction ($SS+OLS^{SCL}$), instead results in a relative RMSE of 0.140. Later, in section **??**, we explain why those results might be too optimistic but argue that they are still valid for relative comparison. Hence, we compare the amount of unexplained variance of the NDVI-based yield prediction. That is $1 - R^2$ (for R^2 values, see appendix table **??**). Consequently, when correcting our NDVI TS, the unexplained variance decreases by:

$$\frac{(1 - 0.705) - (1 - 0.736)}{1 - 0.705} = 10.5\%^{2}$$

Note that the results discussed here depend strongly on the link function used (cf. equation **??**). Once we change it, we should also repeat this analysis.

6.2.2 Investigation of Error Sources in Yield Estimation

Although the YPE was not our primary goal, but was only used as a means to select the best IS, we compare our values with the corresponding ones by **?**. There, a YPE 1.00 [t/ha] was obtained using meteorological data in addition to NDVI TS. Since our error is only about 3.3% larger (cf. table **??**), we consider our results to be competitive. Especially as we did not use meteorological data aside from the timescale transformation (cf. section **??**), and in contrary did not scale the yield down by 10% (cf. section **??**). In the following, we ask ourselves how much modelling performance we can actually expect. This will be limited by multiple sources of uncertainty in the data:

i.) Uncertainty in yield data collected by the combine harvester (**?**).

ii.) Uncertainty in yield data through rasterization.

iii.) Contamination of satellite images through clouds and other atmospheric effects (cf. section **??**).

iv.) Heterogeneity within one pixel that includes, for example, very dense vegetation (and thus according to **?** saturation of the NDVI) on the one half and dry soil at the other half leads to a less informative NDVI value.

[2] The calculation could be also done using the relative RMSE values from table **??** via: $(0.148^2 - 0.140^2)/0.148^2$.

v.) Uncertainty introduced by interpolating NDVI TS, especially when long data-gaps are present.

Even if we had a perfect NDVI curve, it contains only a fraction of the information about the underlying vegetation. Nonetheless, **?** manages to explain up to 86% of the variance in crop yield with only the NDVI TS and meteorological data (Table 5). Although the authors divided the data into training and test data, this subdivision was done randomly at pixel level (without subdividing into fields or years). Thus, there are pixels in the training data that are neighboring pixels from the test data and consequently exhibit high correlations (in yield and spectral reflectances). We suspect that these high values are due to overfitting via high-correlation pixels. This line of argument is consistent with the poor results for cross-year-validation[3] (table 6). The authors, however, account them to uneven (extreme) weather. If this is not rather caused by the suspected overfitting, could be investigated by performing a cross-field-validation[4].Nevertheless, we claim, that our results are not affected by spatial correlation of neighboring pixels. This is because our result is not a 'good' YPE, but the selected IS. Furthermore, we expect all tested ISs to benefit equally from this correlation in terms of YPE, and we are only interested in the relative differences.

6.2.3 NDVI Correction as Unsupervised Learning

The question arises if we can build the correction model on the same year as we want to apply it on. Usually, a similar approach might carry the danger of overfitting. However, we have not used any ground truth at any point (until the evaluation). Instead, we estimated the 'true' NDVI with the assumption **??** via OOB. In other words, we have not used any ground truth but rather developed an unsupervised learner of the NDVI. Consequently, we reason that we can apply our method to a new (comparable) dataset.

6.2.4 Using Additional Covariates

In section **??** we have only used covariates derived from spectral data. We decided against using meteorological data, since we consider five years of data not to be sufficient to model patterns of how vegetation reacts to various weather events. Moreover, we expect the weather in our study region to be rather homogeneous, which is suggested by the fact that the meteorological data published by Meteoswiss are for a grid with a resolution of 1 km. On the other hand, we want the underlying model not to learn improper relationships. For example, the model might automatically predict a high NDVI for a day in summer (detected by high GDD or many sunshine hours) just because it is 'used' to observing a lot of vegetation in summer. Including temporally (e.g., P_{t-1} and P_{t+1}) and geographically adjacent pixels would likely improve performance. However, for simplicity, we omitted it here[5].

[3] By cross-year-validation we understand a cross validation with respect to the RMSE, where each year represents a single fold.

[4] By cross-field-validation we understand the same as with cross-year-validation but with splitting each fold (i.e., a year) further into the respective fields. Since we have multiple fields per year, during evaluation each model trained will have seen the weather of all years but no adjacent pixels.

[5] This is done for simplicity of understanding and using the model, since one would need to adapt to some convention of how to supply the data of adjacent pixels without redundancy (i.e., supplying P_t multiple times). Another complication would be a border-pixel with some adjacent pixels outside the field.

Chapter 7

Conclusion

In this thesis, we investigated how to model vegetation dynamics through NDVI TS derived from satellite images. The major challenges faced, were how to deal with contaminated observations (due to clouds or shadows) and how to interpolate the observed NDVI values. A summary of the IMs considered can be found in the table **??**.

Filtering the observations contaminated by clouds and shadows via SCL introduces data gaps, especially in winter. Therefore, we aim for IMs that handle such data gaps well. The Nadaraya-Watson kernel estimator struggles when there are no or too few points in the window of interest; Universal Kriging is biased towards the mean, particularly in environments with no data (cf. figure **??**); 2^{nd} order Fourier series can deviate strongly within data gaps (cf. figure **??**) and the Savitzky-Golay filter depends on equidistant observations (cf. section **??**). Occasionally, a generalization of the Savitzky-Golay filter — the Locally Weighted Regression — has also shown surprising behavior in data gaps (cf. figure **??**).

In contrast, the latter performed well in Leave-One-Out-Cross-Validation (LOOCV) (cf. table **??**). Nevertheless, we prefer the Smoothing Splines (SS) as they perform only slightly worse there, but produce a much smoother curve (cf. figure **??** and **??**). SS flexibly approximate the data while keeping curvature low (cf. equation **??**). B-splines, on the other hand, were worse than SS with respect to every score function tested, and their smoothing mechanism is also less interpretable. However, the best performing method here is the approximation by a Double logistic (DL), which makes strong assumptions about the shape of the NDVI curve. Problems for the parameter estimation of the DL (and the Fourier series) have been resolved by restricting the parameter space by generous but realistic values. Problems with overfitting in universal kriging were overcome by determining the variogram parameters for a subsample of NDVI TS and finally using the median of each parameter. In the end, we choose DL and SS as our preferred IMs.

The traditional answer to the question of how to deal with contaminated observations is that we only consider observations that are labeled as vegetation or bare soil by the SCL (SCL45). The unreliability of this labeling, however, is illustrated in figure **??**. Moreover, filtered observations (non-SCL45) might still contain valuable information (see section **??**). Therefore, we do not adhere to traditional (SCL) filtration, but instead consider all observations and correct the observed NDVI with uncertainty estimation. For this, we use statistical models that take additional information such as the remaining spectral bands, the current SCL label and the observed NDVI into account. But before we interpolate the

corrected NDVI values, we assign a weight to each observation, corresponding to its uncertainty. The uncertainty is estimated analogously as the NDVI has been corrected. That is, taking the same covariates but replacing the old response ($NDVI^{true}$) with a new one ($abs(NDVI^{corrected} - NDVI^{true})$). By combining different IMs with various statistical models, we obtain 28 different Interpolation Strategies (ISs) (see section **??**). To assess which of these ISs is best, we assume that the better the IS, the better it allows interpolated NDVI TS to predict yield. Surprisingly, the best strategy is the one with SS and the simplest static model considered, which uses only the observed NDVI and SCL classification. Let us recapitulate the best IS: First, we estimate the 'true' NDVI (c.f. assumption **??**) using SS via LOOCV. Then obtain the corrected NDVI using the OLS^{SCL} model (cf. equation **??**). Subsequently, we estimate the absolute error with the OLS^{SCL} model (cf. equation **??**) and thereby obtain weights which are supposed to reflect the reliability of the corrected NDVI (cf. equation **??**). Finally, we perform a weighted interpolation with SS.

To make the IMs more robust to contaminated observations (outliers) that remained after SCL filtration, we generalized an iterative technique. After an initial fit, in each iteration we give less weight to observations with comparatively large residuals and then perform a weighted interpolation (see section **??**). However, after too many iterations, non-contaminated points might get ignored (i.e., given a zero weight). The greatest improvements, on the other hand, were perceived after the first iteration (see figure **??**). For evaluating the generalized robustification technique, we used raw LOOCV performance on the one hand, and the ability to model the NDVI TS for crop yield estimation on the other hand. On the one hand, robustification (narrowly) misses the target of being part of the best IS. On the other hand, we see in table **??** that robustification leads to smaller LOOCV residuals in most cases. That is (except for the Fourier approximation) the QAR^{50} and QAR^{75} are smaller for the robustified ones. Hence, when we expect contaminated observations, we advise to robustify the interpolation.

As to the question of which IM we recommend, we consider two cases. If one only intends to fit a curve to the NDVI TS as precisely as possible, we recommend the robustified DL, since it minimizes the LOOCV residuals in most cases (cf. table **??**). In the event that one requires an interpolation that contains as much information about the plant as possible, we recommend the SS. This recommendation is especially valid if we traditionally consider only SCL45 observations without correcting the proposed NDVI. However, we recommend the abovementioned IS with NDVI correction, because it reduces the unexplained variance of the NDVI-based yield prediction by 5.4% (cf. section **??**). Considering all the error sources (cf. section **??**) and the fact that we only consider the NDVI TS, we consider the 5.4% to be a solid improvement.

7.1 Future Work

7.1.1 Time Series Correction-Interpolation as a General Method

Throughout this thesis, we developed a correction and IM for the NDVI. However, we never relied on any properties of the NDVI. Only the parameter estimated via cross-validation in chapter **??** depends on the scale of the TS. For simplicity, we could thus determine the parameter using Generalized Cross Validation (**?**). Therefore, our approach of interpolation and correction of TS can be applied to arbitrary TS if additional information is available. This includes TS outside of satellite imagery or remote sensing. However, further research is required, to demonstrate the general usefulness of this approach.

As an example, we could develop cloud-correction with uncertainty estimation and interpolation. In the same manner as we corrected the NDVI TS for a pixel in chapter **??**, one could look at each spectral band separately and correct it with an uncertainty estimate. Subsequently, one reassembles the corrected bands and translates the multiple estimated uncertainties into one. Optionally, the TS can also be interpolated before merging, as in chapter **??**. The resulting question would be how well this approach performs.

7.1.2 Minor Improvements

During this project, we also noticed some minor issues that we would have liked to investigate further if more resources were available. The most relevant of these are:

— **Data:** The method how the combine harvester point cloud has been extrapolated to the 10m grid of S2 could possibly be improved.

— **Data:** We have not included the spectral bands that have a resolution of 60 m. But precisely these seem to be promising for cloud correction, since they are a proxy of the water (content and form) in the atmosphere.

— **Data: ?** presents a machine learning approach that supposedly improves the SCL and thus could improve our results that are based on the SCL.

— **NDVI Correction:** Explore the effect of different link and normalizing functions in section **??**. Currently, we run into the danger of some outer points getting nearly ignored just because one estimated absolute residual for some interior point is close to zero.

— **NDVI Correction:** Yield is not the only target variable of interest. Other variables like protein content could also be used in section **??** for the method evaluation.

Appendix A

Reproducibility

A.1 Reproduce Results

For reproducibility of the whole computations, we refer to our codebase at:
https://github.com/LGraz/MasterThesis-Code
In order to reproduce our computations and results, set up the directory as described in the README. The 'Yield Mapping' Data used, is published alongside **?**. Execute the computations via the script `./shell_scripts/reproduce.sh` and do not execute the python and R files by hand (unless you follow the order in `./shell_scripts/reproduce.sh`).

A.2 R–Package

We also provide an `R` package for a general time series correction and interpolation if additional data is available at:
https://github.com/LGraz/CorrectTimeSeries
In our case, we consider the NDVI time series and the additional data consists of the unused spectral bands.

We recommend installing it via the `devtools` package by:
`devtools::install_github("LGraz/CorrectTimeSeries")`

In the following, we shall give a stand-alone example of how the `R` package can be used:

```
library(CorrectTimeSeries)

# load a list of dataframes, each one describes one pixel with the covariates and
      the response
data(timeseries_list)
str(timeseries_list[[1]])

# Train/Load  RF
train_model_myself ← TRUE
if (train_model_myself){
    # Add "true" NDVI (or generally the response), by Out-Of-Bag estimation
    timeseries_list ← lapply(timeseries_list, function(df) {
        df$oob_ndvi ← OOB_est(df$gdd, df$ndvi_observed) # gdd is the time-axis
        df
    })
    # Train correction model
    formula ← "oob_ndvi ~ B02+B03+B04+B05+B06+B07+B08+B8A+B11+B12+scl_class"
    RF ← train_RF_with_fromula(formula, timeseries_list, robustify=TRUE)
} else {
```

```
    data(RF_for_NDVI)
    RF ← RF_for_NDVI
}

# ADD CORRECTION
timeseries_list ← lapply(timeseries_list, function(df) {
    df$corrected_ndvi ← randomForest:::predict.randomForest(RF, df)
    df
})

# Get interpolation for each timeseries
newx ← 1:1000
lapply(timeseries_list, function(df){
    ss ← smoothing_spline(df$gdd, df$corrected_ndvi)
    predict(ss, newx)$y
})
```

Example of how to use the `CorrectTimeSeries` package

Appendix B

Further Material

B.1 Data and Methods

B.1.1 NDVI

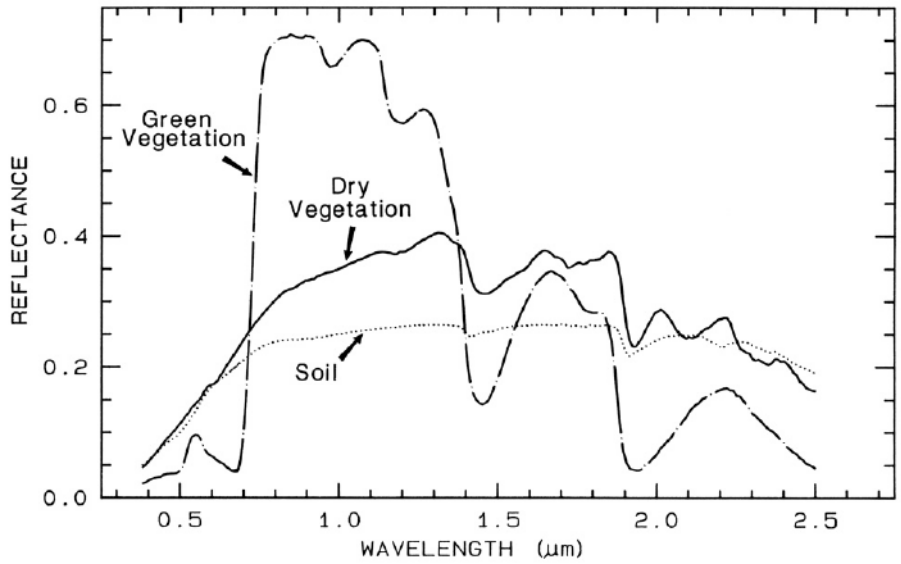


Figure B.1: Spectral reflectance of the green/dry vegetation compared with soil. Note the strong jump at $0.7\mu m$ which is utilized in the NDVI. Figure taken from **?**

B.1.2 GDD

? tabulates the corresponding GDD for each stage of wheat.

Stage	Description	GDD
Emergence	Leaf tip just emerging from above-ground coleoptyle.	125 – 160
Leaf development	Two leaves unfolded.	169 – 208
Tillering	First tiller visible	369 – 421
Stem elongation	First node detectable.	592 – 659
Anthesis	Flowering commences; first anthers of cereals are visible.	807 – 901
Seed fill	Seed fill begins. Caryopsis of cereals watery ripe (first grains have reached half of their final size).	1068 – 1174
Dough stage	Soft dough stage, grain contents soft but dry, fingernail impression does not hold.	1434 – 1556
Maturity complete	Grain is fully mature and drydown begins. Ready for harvest when dry.	1538 – 1665

B.2 Interpolation

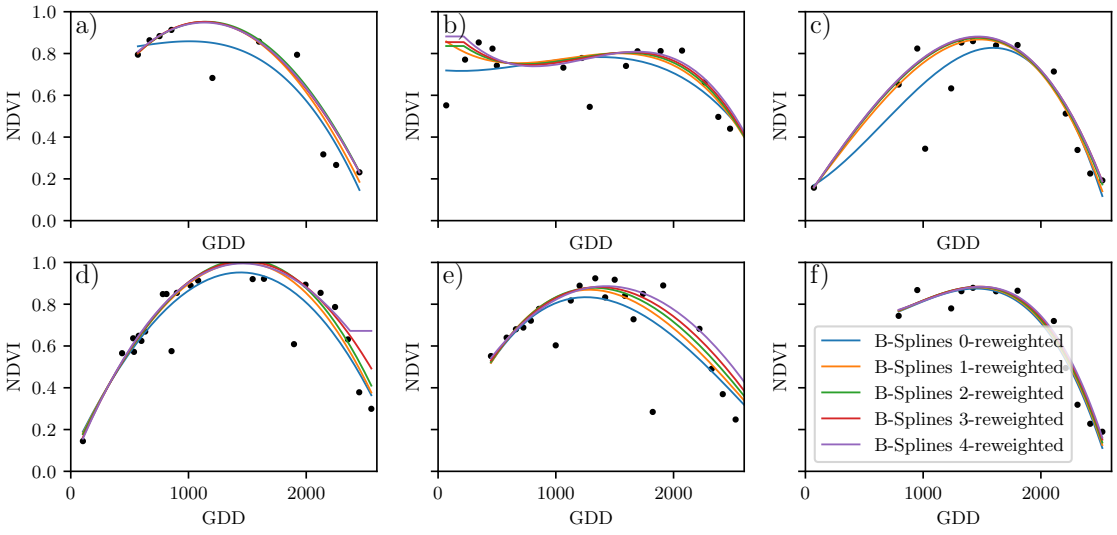


Figure B.2: B-splines fitted to different (SCL45) NDVI TS. Iterations of a robustifying refit (as indicated in section **??**) are also displayed.

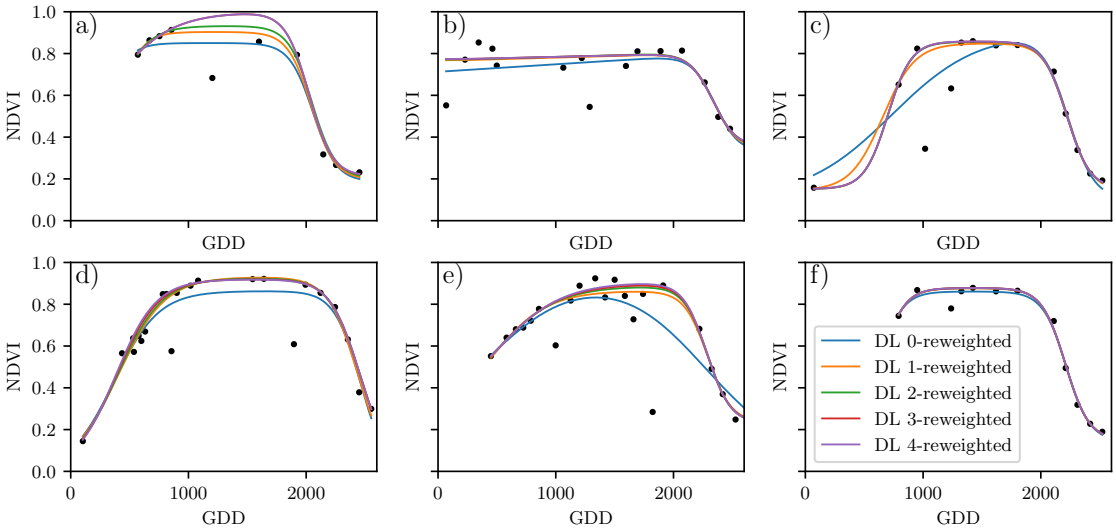
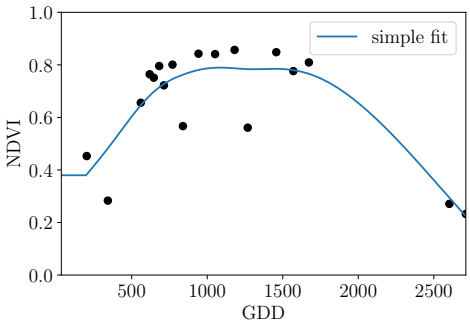
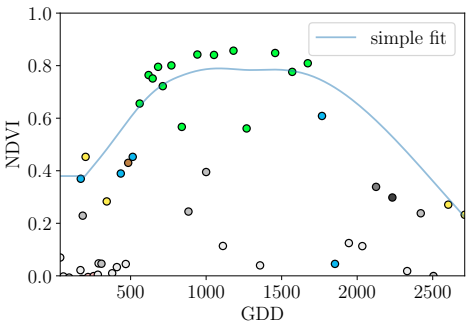


Figure B.3: A Double Logistic curve fitted to different (SCL45) NDVI TS. Iterations of a robustifying refit (as indicated in section **??**) are also displayed.

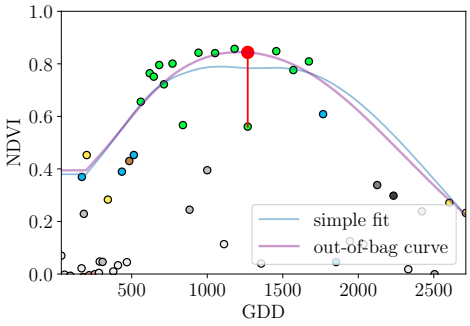
B.3 NDVI correction



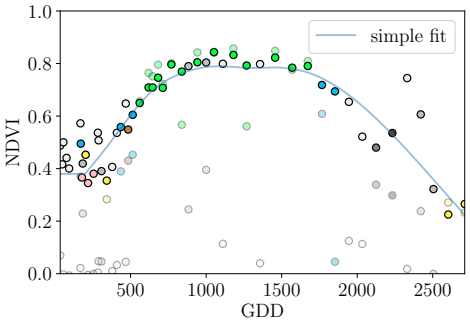
(a) Interpolation via SS (only SCL45)



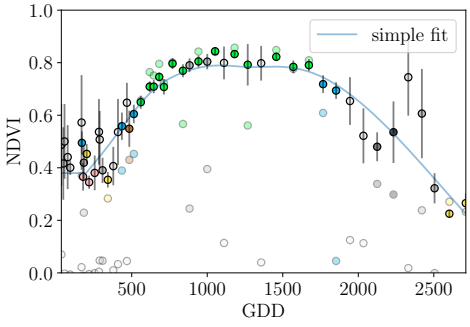
(b) Now also consider other SCL-classes



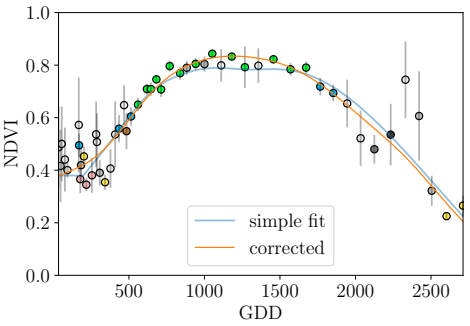
(c) OOB estimation for each point using SCL45



(d) Correct NDVI



(e) Estimate absolute errors via statistical model



(f) Robust interpolation on weights derived from uncertainties.

Figure B.4: Stepwise illustration of robust NDVI-Correction. For the color encoding of the SCL classes we refer to table **??**.

Table B.1: Non-relative RMSE for yield prediction in [t/ha] (cf. table **??**)

	RF	OLS^{SCL}	OLS^{all}	MARS	GAM	LASSO	no corrections
SS	1.144	1.033	1.051	1.042	1.046	1.042	1.095
SS^{rob}	1.144	1.054	1.084	1.094	1.072	1.071	1.091
DL	1.150	1.115	1.116	1.116	1.097	1.098	1.159
DL^{rob}	1.159	1.128	1.117	1.064	1.093	1.105	1.156

Table B.2: Coefficient of determination (R^2) of yield prediction (cf. table **??**)

	RF	OLS^{SCL}	OLS^{all}	MARS	GAM	LASSO	no corrections
SS	0.431	0.486	0.477	0.481	0.479	0.481	0.455
SS^{rob}	0.431	0.475	0.461	0.456	0.467	0.467	0.457
DL	0.427	0.445	0.444	0.444	0.454	0.453	0.423
DL^{rob}	0.423	0.439	0.444	0.470	0.456	0.450	0.424

B.3.1 OLS^{SCL} Model Outputs

```
Call:
lm(formula = (paste(response, " ~ ", "ndvi_observed + scl_class")),
    data = ndvi_df)

Residuals:
    Min      1Q  Median      3Q     Max
-0.7997 -0.0717  0.0039  0.0695  0.6632

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.21465    0.00230   93.46  < 2e-16 ***
ndvi_observed  0.71116    0.00346  205.65  < 2e-16 ***
scl_class3     0.02205    0.00356    6.20  5.8e-10 ***
scl_class4    -0.00431    0.00251   -1.72    0.085 .
scl_class5    -0.09875    0.00234  -42.15  < 2e-16 ***
scl_class6    -0.05301    0.01104   -4.80  1.6e-06 ***
scl_class7     0.11245    0.00274   41.09  < 2e-16 ***
scl_class8     0.25963    0.00253  102.57  < 2e-16 ***
scl_class9     0.35994    0.00236  152.47  < 2e-16 ***
scl_class10    0.09091    0.00308   29.54  < 2e-16 ***
scl_class11    0.29784    0.00392   76.06  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.146 on 124978 degrees of freedom
Multiple R-squared:  0.532,     Adjusted R-squared:  0.532
F-statistic: 1.42e+04 on 10 and 124978 DF,  p-value: <2e-16
```

R Summary of the NDVI correction model (cf. equation **??**)

```
Call:
lm(formula = (paste(get_res(), " ~ ", "ndvi_observed + scl_class")),
    data = ndvi_df)

Residuals:
    Min      1Q  Median      3Q     Max
-0.2051 -0.0427 -0.0074  0.0329  0.6589

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.18647    0.00126  147.74  < 2e-16 ***
ndvi_observed -0.13265    0.00190  -69.80  < 2e-16 ***
scl_class3    -0.00180    0.00196   -0.92   0.3587
scl_class4    -0.04069    0.00138  -29.55  < 2e-16 ***
scl_class5    -0.09698    0.00129  -75.32  < 2e-16 ***
scl_class6    -0.01906    0.00606   -3.14   0.0017 **
scl_class7     0.01641    0.00150   10.91  < 2e-16 ***
scl_class8    -0.00560    0.00139   -4.02  5.7e-05 ***
scl_class9    -0.01384    0.00130  -10.67  < 2e-16 ***
scl_class10   -0.00690    0.00169   -4.08  4.5e-05 ***
scl_class11   -0.01446    0.00215   -6.72  1.8e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.08 on 124978 degrees of freedom
```

```
Multiple R-squared:  0.352,     Adjusted R-squared:  0.352
F-statistic: 6.8e+03 on 10 and 124978 DF,  p-value: <2e-16
```

R Summary of the NDVI correction model (cf. equation **??**)

Declaration of Originality

The signed declaration of originality is a component of every semester paper, Bachelor's thesis, Master's thesis and any other degree paper undertaken during the course of studies, including the respective electronic versions.

Lecturers may also require a declaration of originality for other written papers compiled for their courses.

I hereby confirm that I am the sole author of the written work here enclosed and that I have compiled it in my own words. Parts excepted are corrections of form and content by the supervisor .

Title of work (in block letters):

INTERPOLATION AND CORRECTION OF MULTISPECTRAL SATELLITE IMAGE TIME SERIES

Authored by (in block letters):
For papers written by groups the names of all authors are required.

Name(s):	**First name(s):**
GRAZ	LUKAS

With my signature I confirm that

- I have committed none of the forms of plagiarism described in the Citation etiquette information sheet.
- I have documented all methods, data and processes truthfully.
- I have not manipulated any data.
- I have mentioned all persons who were significant facilitators of the work .
- I am aware that the work may be screened electronically for plagiarism.
- I have understood and followed the guidelines in the document *Scientific Works in Mathematics.*

Place, date:	**Signature(s):**
18.09.2022 Zürich	Graz

For papers written by groups the names of all authors are required. Their signatures collectively guarantee the entire content of the written paper.