



Swiss Federal Institute of Technology Zurich

Seminar for  
Statistics

1   **Department of Mathematics**

2

3

---

4

5   Master Thesis

Spring 2022

6

---

7

**Lukas Graz**

8

9                   **Interpolation and Correction**

10                  of

11                 **Multispectral Satellite Image Time Series**

11

---

12

Submission Date: September 18th 2022

13

---

14

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

# 15 Preface

## 16 Supplementary Material

17 Instructions and the relevant code needed to reproduce this thesis can be found in the  
18 [GitHub repository](#) and to use our results we recommend the provided [R-package](#).  
19 More information is given in the appendix [A](#).

## 20 Acknowledgements

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

24 This endeavor would not have been possible without Gregor Perich. His high personal com-  
25 mitment, reliability as well as the weekly instructive supervision meetings were, without  
26 question, essential for this work.

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

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

# 35 Abstract

36 Multispectral satellite imagery Time Series (TS) are utilized to estimate TS of spectral  
37 indices at the ground. As such, the TS of the Normalized Difference Vegetation Index  
38 (NDVI) is used to model vegetation development. Due to atmospheric effects (e.g., clouds  
39 or shadows) satellite measurements may not match the ground signal. Therefore, traditional  
40 approaches try to filter out contaminated observations before extracting and subsequently  
41 interpolating the NDVI. After filtering, remaining contaminated observations and  
42 resulting data gaps are the two challenges for interpolation that we address in this thesis.

43 For this purpose, we use crop yield maps from 2017-2021 of cereals from a farm in Switzerland  
44 and corresponding Sentinel 2 satellite image TS published by the European Space  
45 Agency. Contaminated observations can be filtered with the provided Scene Classification  
46 Layer (SCL).

47 We give a benchmark-supported review of different interpolation methods and opt for  
48 Smoothing Splines as a flexible non-parametric method and Double Logistic approximation  
49 as a parametric method with implicit shape assumptions. In addition, we generalize an  
50 iterative technique which robustifies interpolation methods against outliers by reducing  
51 their weight. In most cases, this robustification successfully decreased the 50% and 75%  
52 quantiles of the absolute out-of-bag residuals.

53 Moreover, we present a general interpolation procedure that utilizes additional information  
54 to correct the target variable with an uncertainty estimate and then performs a weighted  
55 interpolation. In our setting, the target variable is the NDVI and as additional information  
56 we use the SCL, the observed NDVI and the spectral bands. Consequently, we do not filter  
57 using the SCL but weight observations according to their reliability. The combination of  
58 different interpolation methods and correction models yields 28 interpolation strategies.  
59 In order to choose the best one, we assume that the better the interpolated NDVI TS  
60 models crop growth, the more suitable it is to predict crop yield. Applying this procedure,  
61 the variance in crop yield explained by the resulting NDVI TS decreases by 5.4%.

62 Instructions and a codebase for reproducibility of the results, as well as an R package  
63 making the presented general interpolation procedure accessible to the user, are supplied.

**64 Contents**

65	<b>Notation</b>	<b>v</b>
66	<b>1 Introduction</b>	<b>1</b>
67	<b>2 Data and Methods</b>	<b>3</b>
68	2.1 Sentinel 2 Data . . . . .	3
69	2.2 Crop Yield Data . . . . .	3
70	2.3 Normalized Difference Vegetation Index (NDVI) . . . . .	4
71	2.4 Timescale Transformation . . . . .	5
72	2.5 The Concept of a ‘Pixel’ . . . . .	6
73	2.6 Challenges in S2 Data . . . . .	6
74	2.7 General Methods . . . . .	6
75	2.7.1 Root Mean Square Error (RMSE) . . . . .	8
76	2.7.2 Out-Of-Bag (OOB) and Leave-One-Out-Cross-Validation (LOOCV) . . . . .	8
77	<b>3 Interpolation Methods (IMs)</b>	<b>9</b>
78	3.1 Interpolation Setup . . . . .	9
79	3.2 Parametric Regression . . . . .	9
80	3.2.1 Double Logistic (DL) . . . . .	11
81	3.2.2 Fourier Series (FS) . . . . .	11
82	3.2.3 Optimization Issues . . . . .	12
83	3.3 Non-Parametric Regression . . . . .	12
84	3.3.1 Kernel Regression: Nadaraya-Watson (NW) . . . . .	12
85	3.3.2 Universal Kriging (UK) . . . . .	13
86	3.3.3 Savitzky-Golay Filter (SG) . . . . .	15
87	3.3.4 Locally Weighted Regression (LOESS) . . . . .	16
88	3.3.5 B-Splines (BS) . . . . .	17
89	3.3.6 Smoothing Splines (SS) . . . . .	17
90	3.4 Tuning Parameter Estimation . . . . .	18
91	3.5 Robustification . . . . .	19
92	3.5.1 Our Adjustment: . . . . .	20
93	3.5.2 Examples and Conclusions . . . . .	20
94	3.5.3 Upper Envelope Approach - Penalty for Negative Residuals . . . . .	20
95	3.6 Performance Assessment . . . . .	21
96	<b>4 NDVI Correction</b>	<b>22</b>
97	4.1 Considering other SCL Classes . . . . .	22
98	4.2 Correction Models . . . . .	23
99	4.2.1 Ordinary Least Squares (OLS) . . . . .	23
100	4.2.2 Least Absolute Shrinkage and Selection Operator (LASSO) . . . . .	24
101	4.2.3 General Additive Model (GAM) . . . . .	25
102	4.2.4 Random Forest (RF) . . . . .	25
103	4.2.5 Multivariate Adaptive Regression Splines (MARS) . . . . .	26
104	4.3 Weighted Interpolation . . . . .	27
105	4.4 Resulting Interpolation Strategies (ISs) . . . . .	27
106	4.5 Evaluation via (relative) Yield Prediction Error (relative YPE) . . . . .	28

107	<b>5 Results</b>	<b>31</b>
108	5.1 Goodness of Fit for Selected IMs . . . . .	31
109	5.2 YPE for Tested ISs . . . . .	31
110	<b>6 Discussion</b>	<b>33</b>
111	6.1 IMs . . . . .	33
112	6.1.1 Data Gaps in Time Series . . . . .	33
113	6.1.2 Preselection . . . . .	34
114	6.1.3 Candidate Selection . . . . .	34
115	6.2 NDVI Correction . . . . .	34
116	6.2.1 Choose IS . . . . .	34
117	6.2.2 Investigation of Error Sources in Yield Estimation . . . . .	35
118	6.2.3 NDVI Correction as Unsupervised Learning . . . . .	35
119	6.2.4 Using Additional Covariates . . . . .	36
120	<b>7 Conclusion</b>	<b>37</b>
121	7.1 Future Work . . . . .	39
122	7.1.1 Time Series Correction-Interpolation as a General Method . . . . .	39
123	7.1.2 Minor Improvements . . . . .	39
124	<b>Bibliography</b>	<b>40</b>
125	<b>A Reproducibility</b>	<b>43</b>
126	A.1 Reproduce Results . . . . .	43
127	A.2 R-Package . . . . .	43
128	<b>B Further Material</b>	<b>45</b>
129	B.1 Data and Methods . . . . .	45
130	B.1.1 GDD . . . . .	45
131	B.2 Interpolation . . . . .	46
132	B.3 NDVI correction . . . . .	47
133	B.3.1 OLS <sup>SCL</sup> Model Outputs . . . . .	47

# 134 Notations

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

## 138 Variables

$c$	a (vector of) constant(s)
$\lambda \in \mathbb{R}$	a scalar
$n \in \mathbb{N}$	sample size
$i, j$	indices in $\{1, \dots, n\}$
$n \in \mathbb{R}^n$	time, usually in GDD
$w \in \mathbb{R}^n$	a 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 $y - \hat{y}$
$X \in \mathbb{R}^{n \times p}$	the design matrix. Each row corresponds to one observation and each column to one covariate.
$X_{[:,j]}$	the $j$ -th column of $X$
$X_{[i,:]}$	the $i$ -th row of $X$

## 139 Abbreviations and Objects

NDVI	Normalized Difference Vegetation Index ( <a href="#">Rouse, 1974</a> ).
TS	Time Series.
IM	Interpolation Method. That is a simple <sup>1</sup> method that interpolates data $(t_i, y_i)_{i=1, \dots, n}$ and yields a function $f(t) = y$ , approximating the data.
IS	Interpolation Strategy. This is the category of functions that map $(t_i, y_i)_{i=1, \dots, n}$ to a function $f(t) = y$ , approximating the data. So a IS describes a strategy of how to arrive at an interpolation starting from the data $(t_i, y_i)_{i=1, \dots, n}$ . For this, initial data may be corrected (cf. chapter <a href="#">4</a> ), (possibly different) IMs (iteratively) used, weightings applied (cf. robustification in section <a href="#">3.5</a> ). Note, that strictly speaking every IM also is an IS. But usually we expect an IS to involve a more ‘complex’ procedure.
S2	Sentinel 2 satellites. Two multi-spectral image satellites deployed by the European Space Agency.

---

<sup>1</sup>I.e., no combination of various methods.

---

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 2.2
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 2.1b. 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, $P := \{P_t   t \text{ is a valid sample time within a defined season}\}$
$P^{SCL45}$	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.
DAS	Days After Sowing
GDD	Growing Degree Days – cumulative sum of “max(0, temperature – threshold)”
YPE	(Relative) Yield Prediction Error. See Definition 4.5.0.1
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 2.7.2).
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 2.7.2).

## 140 Statistical Models

DL	Double Logistic (see section 3.2.1)
FS	Fourier Series (see section 3.2.2)
NW	Nadaraya-Watson (see section 3.3.1)
UK	Universal Kriging (see section 3.3.2)
SG	Savitzky-Golay Filter (see section 3.3.3)
LOESS	Locally Weighted Regression (see section 3.3.4)
BS	B-splines (see section 3.3.5)
SS	Smoothing Splines (see section 3.3.6)
OLS	Ordinary Least Squares (see section 4.2.1)
$OLS^{SCL}$	OLS using only the observed NDVI and SCL classes (as factor variables)
$OLS^{\text{all}}$	OLS using the covariates $OLS^{SCL}$ uses and the spectral bands
LASSO	Least Absolute Shrinkage and Selection Operator (see section 4.2.2)
GAM	General Additive Model (see section 4.2.3)
RF	Random Forest (see section 4.2.4)
MARS	Multivariate Adaptive Regression Splines (see section 4.2.5)

141 **Chapter 1**

142 **Introduction**

143 Remote sensing aims to measure target variables efficiently from a distance. In this context,  
144 satellite imagery Time Series (TS) such as the imagery TS of the multi-spectral Sentinel 2  
145 satellites freely distributed by the European Space Agency are used ([ESA, 2022b](#)). Large  
146 scale monitoring of forest and agricultural vegetation dynamics is of great interest to  
147 authorities, insurance companies and environmental and climate researchers. Examples  
148 include crop classification for subsidizing farmers ([Henits et al., 2022](#)) and the creation of  
149 crop models for estimating crop yields or nitrogen concentrations ([Courault et al., 2021](#);  
150 [Perich et al., 2021](#)). In order to transform the high dimensional satellite images into  
151 easily interpretable metrics, spectral indices such as the Normalized Difference Vegetation  
152 Index (NDVI) are used ([Rouse, 1974](#)). The NDVI serves as a proxy for vegetation density  
153 (or chlorophyll content to be more precise), and thus the corresponding TS reflects the  
154 vegetation development. The quality of a satellite image however depends on atmospheric  
155 conditions and thus in case of a dense cloud cover the information content derived from  
156 the NDVI is impaired. Therefore, the European Space Agency also provides a Scene  
157 Classification Layer (SCL), which provides additional metadata about what is observed  
158 (e.g., shadows, clouds, vegetation, etc.) ([ESA, 2022a](#)). So when extracting the NDVI  
159 TS from the Sentinel 2 satellite imagery TS, we can filter out the corrupted observations  
160 using the SCL classification. However, due to this filtration it may occur that we have  
161 no observations for several weeks, especially in winter, or that some observations are  
162 wrongly classified by the SCL (e.g., as vegetation) and thus result in an erroneous NDVI.  
163 Consequently, the main challenge is to interpolate an NDVI TS, which can contain both  
164 large data gaps and outliers.

165 There are several approaches to adequately address this issue. One is to look at the  
166 observed evolution of vegetation density and assume its bell shape for the NDVI TS given  
167 the strong correlation between NDVI and vegetation density. Approaches to model this  
168 include a 2nd order Fourier approximation ([Stöckli and Vidale, 2004](#)) or a Double Logistic  
169 function ([Beck et al., 2006](#)). On the other hand, assumptions are made about more abstract  
170 properties of the curve, such as smoothness or the like. We divide these into local and  
171 global approaches. Nadaraya-Watson ([Strbac et al., 2017](#)), Savitzky-Golay Filter ([Chen  
et al., 2004](#)) and Locally Reweighted Regression ([Omori et al., 2021](#)) use a sliding window  
172 to interpolate the TS stepwise. Global methods like B-Splines ([Gurung et al., 2009](#)) and  
173 Smoothing Splines ([Cai et al., 2017](#)) reduce the squares of all residuals simultaneously,  
174 and Universal Kriging fits a Gaussian process to the data ([Chandola and Vatsavai, 2010](#)).

176 In this thesis, we will discuss strengths and weaknesses of these Interpolation Methods  
177 (IMs) and evaluate them with respect to NDVI interpolation. For this purpose, we use  
178 the Sentinel 2 satellite image TS and crop yield maps of different fields of different wheat  
179 species on a farm in Witzwil, Switzerland over the years 2017-2021. To improve IMs, we  
180 generalize and test an iterative technique that makes interpolations more robust to outliers  
181 by weighting them less. Additionally, we determine how data gaps affect the different IMs.  
182 Furthermore, using NDVI as an example, we present a general interpolation procedure that  
183 utilizes additional information to correct the target variable with an uncertainty estimate  
184 and then interpolates. Thus, we no longer have to filter the observations a priori via the  
185 SCL, but instead correct the observed NDVI and weight the observations via the estimated  
186 uncertainties. Combining IMs with the NDVI correcting models produces 28 Interpolation  
187 Strategies (ISs). We benchmark these against an objective quality measure, which assumes  
188 that the better an NDVI TS models crop growth, the more appropriate it is for estimating  
189 crop yield.

190 The research questions, which are pursued in this thesis, are:

- 191 i.) Which IMs are used in the context of NDVI and what are their advantages and  
192 disadvantages?
- 193 ii.) How may contaminated data be dealt with?
- 194 iii.) How do data gaps affect interpolation?
- 195 iv.) How to deal with data gaps?
- 196 v.) How can we recognize a good interpolation of the NDVI?

197 The thesis is structured as follows: After presenting the available data, illustrating chal-  
198 lenges and defining different concepts in chapter 2, we turn to the two main blocks of  
199 this thesis. On the first, in section 3 we study parametric and non-parametric IMs (ques-  
200 tion i.), generalize an iterative robustification technique (question ii.), and show a way to  
201 evaluate interpolations with out-of-bag residuals (question v.). In section 6.1.1 we discuss  
202 how different IMs respond to data gaps (question iii.), and in section 6.1.2 we preselect  
203 IMs. We evaluate this preselection in 5.1 and select two candidates from different IMs in  
204 section 6.1.3. For the second, we attempt to correct contaminated data with statistical  
205 models in section 4 (question ii.) and utilize previously ignored observations, which we  
206 hope will further reduce data gaps (question iv.). In addition, we compare different ISs  
207 using a vegetation-oriented quality measure (question v.) and describe the results in sec-  
208 tion 5.2. Based on these results, we argue what the best IS is in section 6.2. In addition,  
209 we justify why our NDVI correction can be understood as unsupervised learning and why  
210 we relied only on satellite imagery and not on meterological data for the NDVI correction.  
211 Our conclusions of this thesis, recommendations, as well as an outlook on future work is  
212 given in chapter 7.

213 **Chapter 2**

214 **Data and Methods**

215 We will start by describing the available data and the challenges associated with it. Our  
216 study region is a farm of over 800ha, which is located in western Switzerland. From Perich  
217 et al. (2022) we acquire satellite image data (section 2.1), yield maps of several cereals  
218 from 2017 to 2021 (section 2.2), and meteorological data (section 2.5). Afterwards, we will  
219 introduce general methods in section 2.7, which will be used in the remaining chapters.

220 **2.1 Sentinel 2 Data**

221 The European Space Agency (ESA, 2022b) freely distributes the high-quality images of  
222 the two Sentinel satellites (S2). Together, both satellites have a revisit time of 5 days at  
223 the Equator and 2-3 days at mid-latitudes. However, in our study region, we only receive  
224 an image every 5 days.

225 The S2 images contain 12 spectral bands with spatial resolutions up to 10 meters (see 2.1).  
226 Bands with a lower resolution (20 and 60 meters) were upscaled to 10 meter resolution using  
227 cubic interpolation (Perich et al., 2022). In order to decrease the effect of atmospheric  
228 conditions like reflections and scattering, bottom-of-atmosphere, radiometric corrected  
229 Level-2A data was used<sup>1</sup>. The European Space Agency also supplies an algorithm (ESA,  
230 2022a) produces Scene Classification Layer (SCL) where for each location the observed  
231 subject is assigned to one of 11 SCL-classes (cf. table 2.2). In this thesis, we will use  
232 this classification to filter out data points, that we believe to be less informative. That are  
233 all observations which SCL-class does not correspond to vegetation or bare soils (classes  
234 4 and 5). For convenience, we define the set SCL45 as the observations that belong to  
235 SCL-class 4 or 5.

236 **2.2 Crop Yield Data**

237 The crop yield data were collected using a combine harvester. Equipped with GPS, the  
238 harvester drives over the fields and continuously estimates the dry crop yield density in  
239  $t/ha$  (see fig. 2.1a). We take the data set derived in Perich et al. (2022), where error-  
240 prone measurement points (such as during a tight curve of the combine harvester) were

<sup>1</sup>According to Perich et al. (2022): “Data prior to March 2018 was only available in the top-of-atmosphere L1C format and was downloaded as such [...] L1C data was processed to L2A product level using the ‘Sen2Cor’ processor provided by the European Space Agency”.

Table 2.1: List of spectral bands of the S2-satellites. Each band has its center at the wavelength  $\lambda$  in nm with the spectral width  $\Delta\lambda$  in nm with a spatial resolution  $SR$  in m (Jaramaz et al., 2013).

Band	$\lambda$	$\Delta\lambda$	$SR$	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
[Black]	0:	Missing Data	[Blue]	6:	Water
[Red]	1:	Saturated or defective pixel	[Dark Gray]	7:	Cloud low probability
[Dark Gray]	2:	Dark features / Shadows	[Light Gray]	8:	Cloud medium probability
[Brown]	3:	Cloud shadows	[Light Blue]	9:	Cloud high probability
[Green]	4:	Vegetation	[Pink]	10:	Thin cirrus cloud
[Yellow]	5:	Bare soils	[Light Red]	11:	Snow or ice

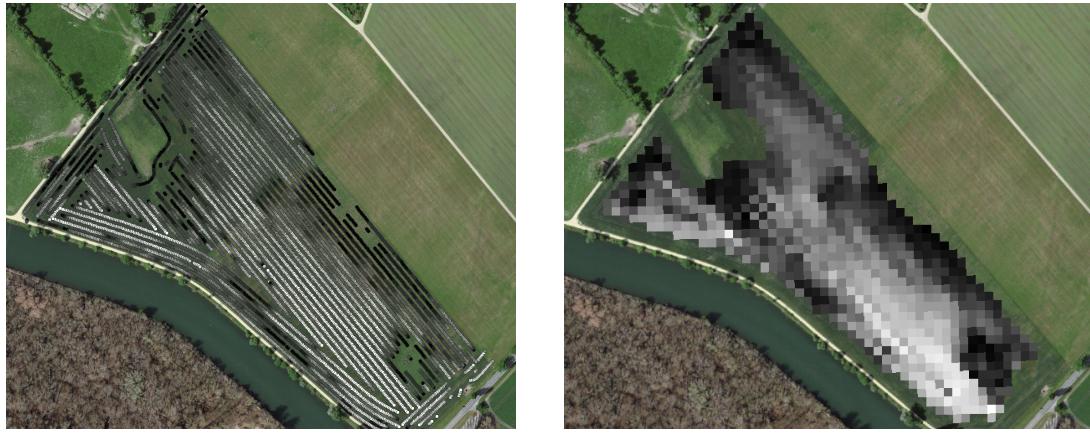
241 removed and then the yield map was rasterized using linear interpolation (cf. fig. 2.1b).  
 242 We summarize the rasterized dry-yield values by the following statistics:

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

244 Comparing the average per-field crop yield reported by the farmer with the yield estimated  
 245 by the combine harvester shows that the latter overestimates crop yield by ca. 10% (Perich  
 246 et al., 2022). Since the relative estimation error is approximately constant and we do not  
 247 aim for an accurate yield prediction, we will not consider this deviation.

## 2.3 Normalized Difference Vegetation Index (NDVI)

249 The well-known (NDVI) introduced in Rouse (1974) is used to measure vegetation in  
 250 remote sensing. It utilizes a large jump of reflectancy between red and infrared and can



(a) Raw combine harvester data (cleaned).

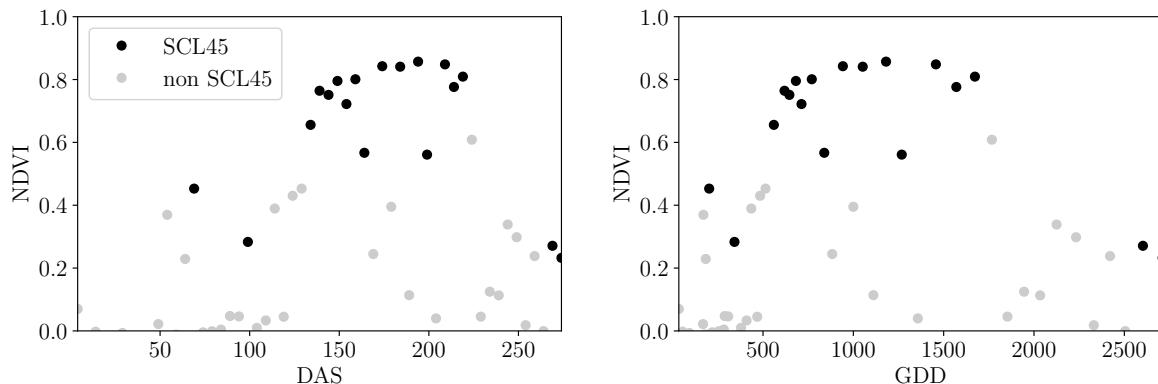
(b) Rasterized to Sentinel 2 resolution.

Figure 2.1: Crop yield density map of a field. Ranges from 0.1 t/ha (black) to 5.35 t/ha (white)

251 be calculated using the bands  $B4$  and  $B8$  (table 2.1) by:

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

252 Since we measure the NDVI via the S2 satellites from space we can not expect to measure  
 253 the true NDVI. This is especially true if we do not see the ground because of clouds or the  
 254 ground signal is disturbed by cloud shadows. Even if we only use SCL45 observations we  
 255 still encounter issues as will be described in section 2.6. Therefore, we call the calculated  
 256 values merely the observed NDVI. In the following chapters, we will study the resulting  
 257 NDVI TS (for one location and one season) extensively. Such a TS is shown in figure 2.2a.



(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.

258

## 259 2.4 Timescale Transformation

260 Regarding the Days After Sowing (DAS) time scale shown in fig. 2.2a, we detect two  
 261 drawbacks. First, this scale makes it difficult to compare two NDVI TS because wheat is  
 262 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 the winter, we face significant data gaps in this period. The time scale transformation introduced in [McMaster and Wilhelm \(1997\)](#) fixes both problems. The resulting Growing Degree Days (GDD) are defined as the cumulative sum since sowing of temperature above a given base temperature  $T_{base}$ . For cereals, we use  $T_{base} = 0$  ([Perich et al., 2022](#)). 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 tabultaed in [B.1.1](#). In figure [2.2](#) 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 succesfully compressed. Due to 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 confront the tasks in the following chapters.

Consider a 10 by 10 meter square that coinsides with a S2 image pixel and  $T$  the GDD values for which S2 images are avialable in a given season. For  $t \in T$  let  $P_t$  be a tupel 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-class 4 or 5 (vegetation and soil).

## 2.6 Challenges in S2 Data

Now, we shall illustrate with an example pixel the challenges, we will confront in the coming chapters. The figure [2.3](#) 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, we see a pale green and we also note a NDVI.

So in conclusion, 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 corrupt ones.

## 2.7 General Methods

Here we will only introduce Methods that will accure at several places. For IMs we refer to sections [3.2](#) and [3.3](#), for a robust IS to section [3.5](#). In section [3.4](#) we describe a method

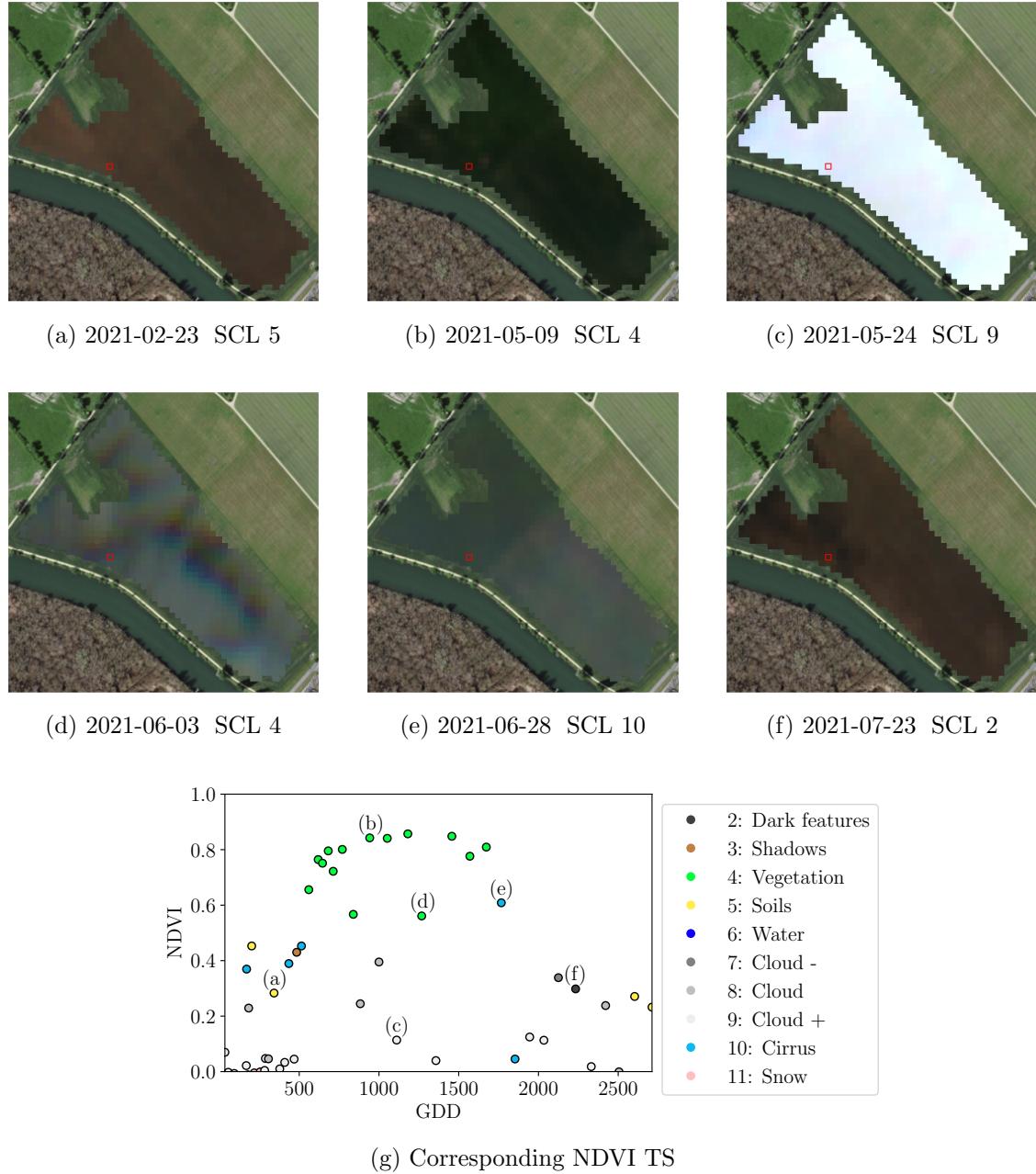


Figure 2.3: Satellite images of a field at selected times with a static background for orientation. Moreover, the NDVI TS of the red-highlighted pixel is shown in (g) colored by the SCL labels.

301 to objectively determine the quality of an interpolation, and in chapter 4 we present the  
 302 NDVI correction together with an adapted IS.

303 **2.7.1 Root Mean Square Error (RMSE)**

304 In this section we describe different criteria to evaluate models. Hence, given a vector  
 305  $y \in \mathbb{R}^n$  and its estimator  $\hat{y}$  (estimated using the model), we define the RMSE as:

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

306 **2.7.2 Out-Of-Bag (OOB) and Leave-One-Out-Cross-Validation (LOOCV)**

307 The rationale for OOB and LOOCV is that we intend to evaluate a model  $M$  with unseen  
 308 data. That is, if  $D$  describes the entire dataset and we train a model on a subset of  $D$ , we  
 309 can use the remaining data to evaluate the model.

To formally introduce this, let:

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

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

313 In the bootstrap (e.g., random forest) framework, we define  $\hat{y}_i$  to be the average of all  
 314 computed and admissible  $M^{(-i)}$ .

315 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  
 316 the corresponding  $\hat{y}_i$  also the LOOCV estimator.

317 If we optimize some parameter via OOB (or LOOCV) this means that we search for the  
 318 parameter that minimizes some loss function which takes the OOB (or LOOCV) residuals.  
 319 Usually we approximate this parameter by searching on a grid.

320 **Chapter 3**

321 **Interpolation Methods (IMs)**

322 In section 2.6 we have established the need for interpolating the NDVI TS. In this chapter  
323 we first specify a setting for the interpolation and divide the IMs into those that  
324 make fundamental shape assumptions (parametric) and those that are more flexible (non-  
325 parametric). We give an introduction for each method with a compact definition, highlight  
326 adjustments or give remarks where appropriate, and then point out strengths and  
327 weaknesses of each method. Additionally, a brief overview of the considered IMs is pro-  
328 vided in table 3.1. Afterwards, we extract an robustification strategy from the one IM and  
329 generalize it so we can use it for all methods that allow for a priori weighted observations.  
330 Finally, using LOOCV, we tune the parameters (where necessary) and get a first idea of  
331 the performance of each method.

332 **3.1 Interpolation Setup**

In this chapter, we will only consider SCL45 observations, since they are more reliably. 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 time  $t_i$ . Assume that it can be represented by

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

where  $\varepsilon_i$  is some noise and  $m : \mathbb{R} \rightarrow \mathbb{R}$  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

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

333 We will introduce parametric and non-parametric approaches to estimate  $m$  in section 3.2  
334 and 3.3 Furthermore, in the subsequent, we denote  $w \in \mathbb{R}^n$  as the vector of weights such  
335 that  $w_i$  corresponds to the weight that  $(t_i, y_i)$  should have in the interpolation.

336 **3.2 Parametric Regression**

337 Parametric Curve estimation tries to fit a parametric function, such as, for example, a  
338 Gaussian function with parameters  $\mu$  and  $\sigma$ , to a dataset. In the following, we introduce  
339 two parametric approaches.

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).

	<b>Assumptions</b>	<b>Advantages</b>	<b>Disadvantages</b>	w	b
Double- Logistic	<ul style="list-style-type: none"> <li>- Function first increases then decreases</li> <li>- NDVI has a minimal value</li> </ul>	<ul style="list-style-type: none"> <li>- Good for evergreen plants (if snow masks NDVI)</li> <li>- Upper envelope</li> </ul>	<ul style="list-style-type: none"> <li>- Parameter estimation can be very difficult</li> <li>- Strange behavior for long data-gaps</li> </ul>	Yes	(Yes)
Fourier Series	<ul style="list-style-type: none"> <li>- NDVI can be approximated by a 2cd order Fourier series.</li> </ul>	<ul style="list-style-type: none"> <li>- Incorporates periodical growth-cycles</li> </ul>	<ul style="list-style-type: none"> <li>- Parameter estimation can be very difficult</li> <li>- Curve easily exceeds bounds of the NDVI</li> </ul>	Yes	No
Nadaraya- Watson (Kernel Smooth- ing)	<ul style="list-style-type: none"> <li>- Close points are related to each other via a kernel function</li> </ul>	<ul style="list-style-type: none"> <li>- Simple</li> <li>- Computationally very fast</li> </ul>	<ul style="list-style-type: none"> <li>- Biased, especially at ‘peaks’ and ‘valleys’</li> <li>- Bandwidth: fails if there are big data-gaps</li> </ul>	Yes	Yes
Universal Kriging	<ul style="list-style-type: none"> <li>- Function is a realization of a stationary Gaussian process</li> </ul>	<ul style="list-style-type: none"> <li>- Informative parameters</li> <li>- Flexible</li> </ul>	<ul style="list-style-type: none"> <li>- Regression to the mean</li> <li>- Assumptions clearly not met</li> </ul>	Yes	(Yes)
SG	<ul style="list-style-type: none"> <li>- High frequencies are noise (Low-Pass-Filter)</li> <li>- Equidistant points</li> <li>- Local polynomials</li> </ul>	<ul style="list-style-type: none"> <li>- Computationally very fast</li> </ul>	<ul style="list-style-type: none"> <li>- Cannot deal natively with missing data (need some interpolation)</li> </ul>	No	(Yes)
SG + NDVI	<ul style="list-style-type: none"> <li>- Upper envelope</li> <li>- Vegetation cannot grow faster than some slope</li> </ul>	<ul style="list-style-type: none"> <li>- Biological knowledge</li> </ul>	<ul style="list-style-type: none"> <li>- Bad “upper envelope” since weights are not used for the estimation itself</li> </ul>	(No)	(Yes)
LOESS	<ul style="list-style-type: none"> <li>- Local polynomial with points closer to the estimated point are more important</li> </ul>	<ul style="list-style-type: none"> <li>- Flexible</li> <li>- Generalization of SG</li> <li>- Weighting function makes intuitive sense</li> </ul>	<ul style="list-style-type: none"> <li>- Computationally expensive</li> </ul>	Yes	(Yes)
B-Splines (Smoothed)	<ul style="list-style-type: none"> <li>- Function can be approximated by a linear combination of B-splines basis functions</li> </ul>	<ul style="list-style-type: none"> <li>- General assumption</li> <li>- Flexible shape</li> </ul>	<ul style="list-style-type: none"> <li>- Unbounded</li> <li>- No intuitive meaning for smoothing</li> </ul>	Yes	No
Smoothing splines	<ul style="list-style-type: none"> <li>- 2cd derivative of function is integrable</li> </ul>	<ul style="list-style-type: none"> <li>- Intuitive meaning of penalty</li> <li>- General assumptions</li> <li>- Flexible shape</li> </ul>	<ul style="list-style-type: none"> <li>- Choice of smoothing parameter</li> </ul>	Yes	(Yes)

340 **3.2.1 Double Logistic (DL)**

The Double Logistic smoothing as described in Beck et al. (2006) 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:

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

- 341 Where the five free parameters:  $y_{\max}$ ,  $d_0$ ,  $d_1$ ,  $t_0$ ,  $t_1$  are initially estimated by least squares.  
 342 Such fit can be seen in figure 3.1.

343 **Robustification**

- 344 Similar as for the SG (cf. section 3.3.3) one can reestimate (only once) the parameters by  
 345 giving less weight to the overestimated observations and more weight to the underestimated  
 346 observations. For the details on the choice of the weights we refer to Beck et al. (2006).  
 347 We will not apply this reestimation but rather the robustification introduced later in  
 348 section 3.5.

Advantages	Disadvantages
<ul style="list-style-type: none"> <li>— Incorporates subject specific knowledge in the case of evergreen plants covered in snow.</li> <li>— Optimized parameters have an intuitive meaning.</li> </ul>	<ul style="list-style-type: none"> <li>— Strong shape assumptions on the NDVI curve.</li> <li>— Parameter optimization might go wrong. This can be mitigated to some extent to provide bounds for the parameters.</li> <li>— Strange behavior in regions with little observations (cf. figure 3.1).</li> </ul>

349 **3.2.2 Fourier Series (FS)**

Stöckli and Vidale (2004) approximates the NDVI curve using a second order FS:

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

- 350 where  $\Phi = 2\pi \times (t - 1)/n$ . Thus, we periodical behavior. If we would set the period to  
 351 match one year this would coinced with the notion that plans grow every year. Analogous  
 352 to section 3.2.1 we fit it to the data by least squares. Example fits can be seen in figure 3.1

Advantages	Disadvantages
<ul style="list-style-type: none"> <li>— Assumption of periodicity can be helpful if we are modelling multiyear growth cycles.</li> <li>— Flexible curve shape.</li> </ul>	<ul style="list-style-type: none"> <li>— Bad behavior in regions with little data (cf. figure 3.1).</li> <li>— Hard to interpret estimated parameters.</li> <li>— Parameter estimation can go wrong. Introducing bounds can help.</li> </ul>

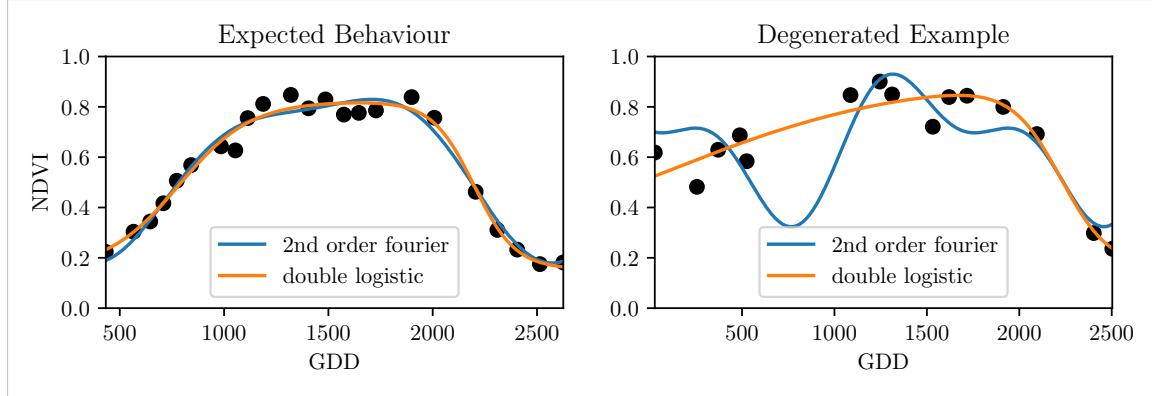


Figure 3.1: Here we observe the possibilities of a precise fit for the two parametric methods but notice also some misbehavior.

### 353 3.2.3 Optimization Issues

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

## 359 3.3 Non-Parametric Regression

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

### 363 3.3.1 Kernel Regression: Nadaraya-Watson (NW)

364 As described in section 3.1, we aim to estimate

$$\mathbb{E}[Y \mid T = t] = \int_{\mathbb{R}} y f_{Y|T}(y \mid t) dy = \frac{\int_{\mathbb{R}} y f_{T,Y}(t,y) dy}{f_T(t)}, \quad (3.3.1.1)$$

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

$$\hat{f}_T(t) = \frac{\sum_{i=1}^n K\left(\frac{t-t_i}{h}\right)}{nh}, \quad \hat{f}_{T,Y}(t,y) = \frac{\sum_{i=1}^n K\left(\frac{t-t_i}{h}\right) K\left(\frac{y-Y_i}{h}\right)}{nh^2}, \quad (3.3.1.2)$$

<sup>1</sup>We used the python function `scipy.optimize.curve_fit`.

where  $h$ , the bandwidth, symbolizes the windowsize of to consider. By using the above function in equation (3.3.1.1) we arrive at the NW kernel estimator:

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

367 Common choices for the kernel are the normal function or a uniform function (also called  
368 ‘bot’ function).

### 369 Choose Bandwidth

370 Note that we still need to choose the bandwidth of the function. This can be done with  
371 the help of LOOCV while optimizing the RMSE. For non-equidistant data we refere to  
372 Brockmann et al. (1993) 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 $\hat{\sigma}_\varepsilon^2$ . <sup>2</sup>	

### 373 3.3.2 Universal Kriging (UK)

374 UK as described in dig (2007) was developed in geostatistics to deal with autocorrelation  
375 of the response variable at locations that are spatially close. By applying the notion that  
376 two spectral indices that are timewise close should also take similar values, we justify the  
377 application of UK. In the end, we would like to fit a smooth Gaussian process to the data.

378 A Gaussian Process  $\{S(t) : t \in \mathbb{R}\}$  is a stochastic process if  $(S(t_1), \dots, S(t_k))$  has a multi-  
379 variate Gaussian distribution for every collection of times  $t_1, \dots, t_k$ .  $S$  can be fully charac-  
380 terized by the mean  $\mu(t) := E[S(t)]$  and its covariance function  $\gamma(t, t') := \text{Cov}(S(t), S(t'))$ .  
381 Furthermore, we will assume the Gaussian process to be stationary. That is for  $\mu(t)$  to be  
382 constant in  $t$  and  $\gamma(t, t')$  to depend only on  $h = t - t'$ . Thus, we will write in the following  
383 only  $\gamma(h)$ .<sup>3</sup>

Now, we need to make some assumption on the covariance function. For this we introduce the variogram of a Gaussian process as

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

and define  $\gamma$  via the above equation by choosing the Gaussian Variogram defined by

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

384 Here  $h$  denotes the distance,  $n$  is the nugget,  $r$  is the range and  $p$  is the partial sill. The  
385 influence of the parameters is visualized in figure 3.2.<sup>4</sup>

<sup>3</sup>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.

<sup>4</sup>Strictly speaking we use a scaled version of the variogram. Thus, only the ratio of  $p/n$  matters.

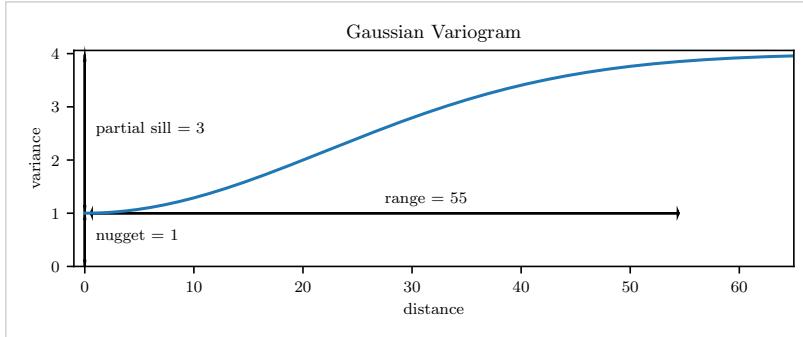


Figure 3.2: Gaussian Variogram with nugget=1, partial sill=3, range=55

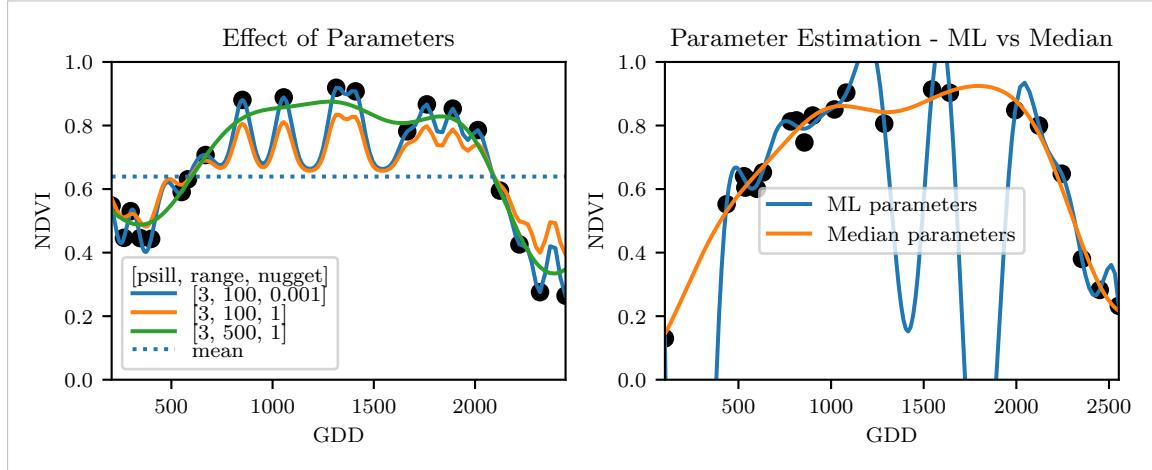


Figure 3.3: On the left, we see how the interpolation change if we increase the nugget and the range parameter. On the right, we compare two UK interpolations, where one takes parameters by numerically maximizing the (which results in a very small nugget) and the other takes the median of many such numerical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

386 Finally, we consider a one-dimensional Gaussian process  $G_\gamma$  with variogram  $\gamma$  and tune the  
 387 variogram parameters using maximum likelihood<sup>5</sup>. Let  $z$  be a vector with the new values  
 388 to extrapolate, then we can determine the values  $m(z) = \mathbb{E}[G_\gamma(z)|(t, y)]$  using Bayes rule<sup>6</sup>.  
 389 For an example fit, we refer to figure 3.3.

### 390 Violated Assumption

391 Since we observe a clear pattern of a growth period in spring and harvest in the end  
 392 of summer, we have to admit that our stationarity assumption with the constant mean  
 393 is structurally violated. This is also the reason why we observe (for every variogram  
 394 parameter) a tendency to the mean, as indicated in figure 3.3.

<sup>5</sup> As illustrated in figure 3.3 maximum likelihood estimation can lead to overfitting. Thus, we will in practice sample several such optimized parameters and use their median in the end.

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

Advantages	Disadvantages
<ul style="list-style-type: none"> <li>— It is a well-studied method.</li> <li>— Variogram parameters have an intuitive meaning.</li> <li>— Flexible covariance structure.</li> </ul>	<ul style="list-style-type: none"> <li>— Regression to the mean.</li> <li>— Violated assumption of constant mean and constant variance. Thus, the NDVI is not a stationary process.</li> <li>— Pure maximum likelihood can result in overfitting.</li> </ul>

### 3.3.3 Savitzky-Golay Filter (SG)

The SG, introduced in [Savitzky and Golay \(1964\)](#) is a technique in signal processing and can be used to filter out high frequencies (low-pass filter) ([Schafer, 2011](#)). Furthermore, 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,  $j \in \{m, m+1, \dots, n-m\}$  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.

[Chen et al. \(2004\)](#) developed a ‘robust’ IM for the NDVI based on 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<sup>7</sup> and by repeating this step we approach the “upper NDVI envelope”.

### 413 Extension: Spatial-Temporal SG

One notable adaptation of the SG is the presented by [Cao et al. \(2018\)](#). The key difference is the additional assumption of the cloud cover being discontinuous and that we can

<sup>7</sup>Here we call a point  $i$  an outlier if  $X_i^0 < X_i^1$ .

416 improve by looking at adjacent pixels<sup>8</sup>. Because we are working with rather high resolution  
 417 satellite data, and we need the variance in the predictors, we will waive this extension.

Advantages	Disadvantages
<ul style="list-style-type: none"> <li>— Popular technique in signal processing.</li> <li>— Efficient calculation for equidistant points.</li> <li>— Upper envelope matches intuition for the NDVI. Therefore, it is robust against outliers with small values.</li> </ul>	<ul style="list-style-type: none"> <li>— No natural way of how to estimate points that are not in the data.</li> <li>— Not generalizable to other spectral indices.</li> <li>— Linear interpolation to account for missing data might be not appropriate.</li> <li>— No smooth interpolation between two measurements.</li> </ul>

418 **3.3.4 Locally Weighted Regression (LOESS)**

419 The LOESS introduced by [Cleveland \(1979\)](#) can be understood as a generalization of the  
 420 SG (cf. sec. 3.3.3).

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

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

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

423 **Differences between the Robust LOESS and the SG**

424 The LOESS smoother takes a fraction of points instead of a fixed number and therefore  
 425 automatically adapts to the size of the data we wish to interpolate. However, we run  
 426 into the danger of considering too little observations, since the estimation breaks down if  
 427  $\lceil \alpha n \rceil < d + 1$ .<sup>9</sup> Furthermore, LOESS gives less weight to points further away. This yields a  
 428 "smoother" estimate, since when we slide the window (e.g., for estimating the next value)  
 429 an influential point at the border does not suddenly get zero weight from being weighted  
 430 equally before. Finally, the LOESS also can be used for non-equidistant data and allows  
 431 for arbitrary interpolation.

Advantages	Disadvantages
<ul style="list-style-type: none"> <li>— Flexible generalization of SG.</li> <li>— Arbitrary interpolation possible.</li> <li>— Intuitive parameters.</li> </ul>	<ul style="list-style-type: none"> <li>— The nature of local regression might lead to surprising estimates (no smoothness guarantees for the second derivative).</li> </ul>

<sup>8</sup>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.

<sup>9</sup>If too many weights are set to zero, we might end up considering not enough observations and thus get a singular design-matrit (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  $\alpha$  is big enough.

432 **3.3.5 B-Splines (BS)**

BS as discussed in [Lyche and Mørken \(2005\)](#) are piecewise cubic polynomials defined by

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

433 where  $B$  are basis functions and recursively defined by:

$$434 \quad \begin{aligned} B_{i,0}(z) &= 1, \text{ if } t_i \leq z < t_{i+1}, \text{ otherwise } 0 \\ B_{i,k}(z) &= \frac{z - t_i}{t_{i+k} - t_i} B_{i,k-1}(z) + \frac{t_{i+k+1} - z}{t_{i+k+1} - t_{i+1}} B_{i+1,k-1}(z). \end{aligned}$$

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 that we have to perfectly interpolate. Thus, we use the minimum number of basis functions<sup>10</sup> such that:

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

Advantages	Disadvantages
— Can be assigned degrees of freedom.	— Smoothing process does not translate well to a interpretation (unlike SS).
— Extendable to "smooth" version.	— Choice of smoothing parameter $s$ .
— Performs also well if points are not equidistant.	

435 **3.3.6 Smoothing Splines (SS)**436 Let  $\mathcal{F}$  be the Sobolev space (the space of functions of which the second derivative is  
437 integrable). Then the unique<sup>11</sup> minimizer

$$\hat{m} := \arg \min_{f \in \mathcal{F}} \sum_{i=1}^n w_i (y_i - f(t_i))^2 + \lambda \int f''(t)^2 dt \quad (3.3.6.1)$$

438 is a cubic spline (i.e., a piecewise cubic polynomial function). The objective function  
439 ensures that we decrease the curvature while keeping the RMSE low.440 **Whittaker — Discrete Version with Higher Order Derivatives**

The Whittaker smoother introduced in [Eilers \(2003\)](#) is closely reminiscent of the SS and is also used for the NDVI TS ([Atzberger and Eilers, 2011](#)). Similar to SS, we minimize the following expression over  $z \in \mathbb{R}^n$ :

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

<sup>10</sup>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.

<sup>11</sup>Strictly speaking it is only unique for  $\lambda > 0$ .

441 where  $W$  is a diagonal weight-matrix,  $\lambda$  our parameter and  $D$  a matrix that serves the  
 442 purpose of approximating a differentiation of  $k$ -th order. In essence this minimization  
 443 function is the same as equation 3.3.6.1. The only difference are, that we substitute  
 444 the integral by a sum and that we are more flexible with the order of the derivatives we  
 445 are using. The main drawback is that we do not get a smooth function that interpolates  
 446 and that the sum behaves worse than the integral for non-equidistant datapoints. Thus,  
 447 we will not consider the Whittaker further but consider the more general SS.

Advantages	Disadvantages
<ul style="list-style-type: none"> <li>— Can be assigned degrees of freedom (trace of the hat-matrix).</li> <li>— Efficient estimation (closed form solution).</li> <li>— Intuitive penalty (we don't want the function to be too "wobbly" — change slopes).</li> <li>— Also performs well if points are not equidistant.</li> <li>— Fixes the Runge's phenomenon (fluctuation of high degree polynomial interpolation).</li> <li>— Bounded within the datarange if <math>\lambda</math> is choosen a priori.</li> </ul>	<ul style="list-style-type: none"> <li>— The tuning parameter <math>\lambda</math> must be chosen. This can be done via cross validation and optimizing a score function (e.g., the RMSE).</li> </ul>

## 448 3.4 Tuning Parameter Estimation

449 Many of the IMs introduced in section 3.2 and 3.3 include a free parameter. To determine  
 450 this parameter for a specific IM, we will estimate the absolute residuals using OOB esti-  
 451 mation and then optimize the parameter using a score function. We clarify the procedure  
 452 step by step:

- 453 i.) Construct a set  $\Lambda$  of candidate parameters that generously covers the parameter  
 454 space.
- 455 ii.) Consider  $\mathcal{P}$ , a set of Pixels.
- 456 iii.) For each parameter  $\lambda \in \Lambda$  consider the individual pixels and compute the LOOCV<sup>12</sup>  
 457 for the absolute residuals of the specific NDVI IM for all Pixels in  $\mathcal{P}$  and store them  
 458 in the set  $R_\lambda$ .
- 459 iv.) Determine  $\lambda_{optimal} = \arg \min_{\lambda \in \Lambda} q_{90}(R_\lambda)$ , where we describe the 90% quantile with  
 460  $q_{90}$ .

461 We choose quantile(90) as our optimization function because we want to allow 10% of  
 462 outliers (corrupt points) but also aim for an accurate fit in 90% of the cases.

463 Figure 3.4 exemplifies the effect of the optimization function (different quantiles). To  
 464 summarize, we may say that the higher the quantile, the stronger the smoothing.

<sup>12</sup>For a definition of the leave-one-out-cross-validation we refer to section 2.7.2.

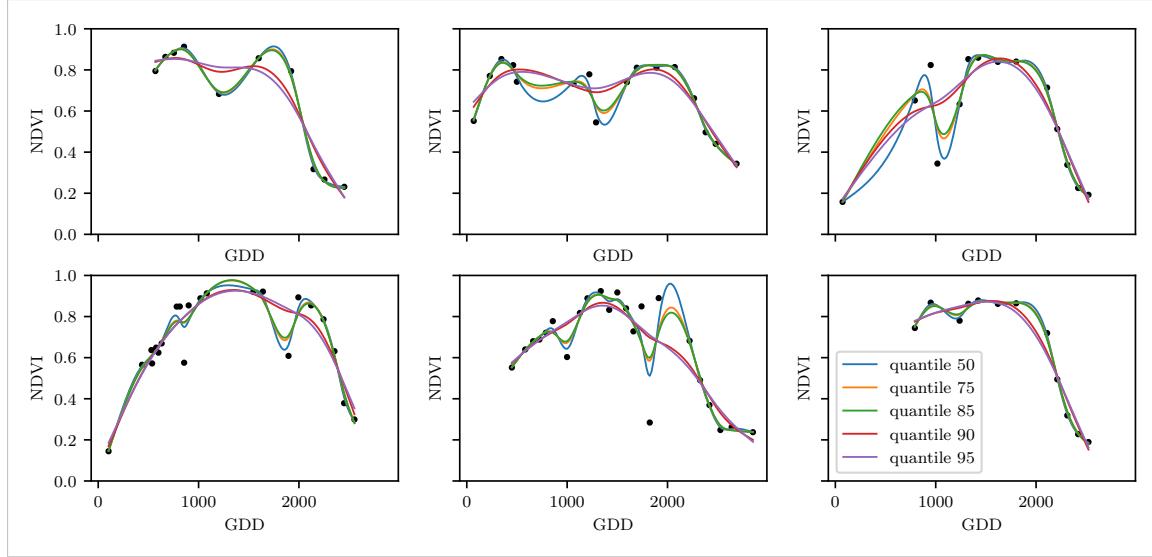


Figure 3.4: Smoothing splines fit with smoothing parameter optimized by minimizing the given quantile of the absolute leave-one-out residuals. Note that the larger the considered quantile is, the smoother the resulting curve becomes.

## 465 3.5 Robustification

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

469 Even though the procedure is taken from the robust version of the LOESS smoother (cf.  
 470 section 3.3.4 and Cleveland (1979)), we can apply it to every IM that allows for prior  
 471 weighting of observations.

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

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

474 Next, we compute new weights by

$$w_i^{\text{new}} := w_i^{\text{old}} \begin{cases} (1 - \tilde{r}_i^2)^2, & \text{if } |\tilde{r}_i| < 1 \\ 0, & \text{else} \end{cases}; \quad (3.5.0.1)$$

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

477 Note that this procedure is indeed robust since we use the median for the normalization  
 478 which has a breakdown point<sup>13</sup> of 50%.<sup>14</sup>

<sup>13</sup>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%.

<sup>14</sup>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).

479 **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

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

480 for  $r, w \in \mathbb{R}^n$ .481 **3.5.2 Examples and Conclusions**

482 Examples of the first four iterative fits using SS are shown in figure 3.5 for six pixels.  
 483 For the analogous figures of the other IMs cf. figures B.1, B.2, B.3 and B.1. Indeed, we  
 484 observe how the interpolated TS is less affected by outliers after each iteration. We notice  
 485 the biggest difference in the first iteration. Furthermore, in the plot at the bottom left we  
 486 see how the interpolation ‘escapes’ from the right endpoint with each successive iteration,  
 487 even though our intuition does not necessarily identify this point as an outlier. Therefore,  
 488 in the following, we will always stop after one iteration.

consider naming the subplots

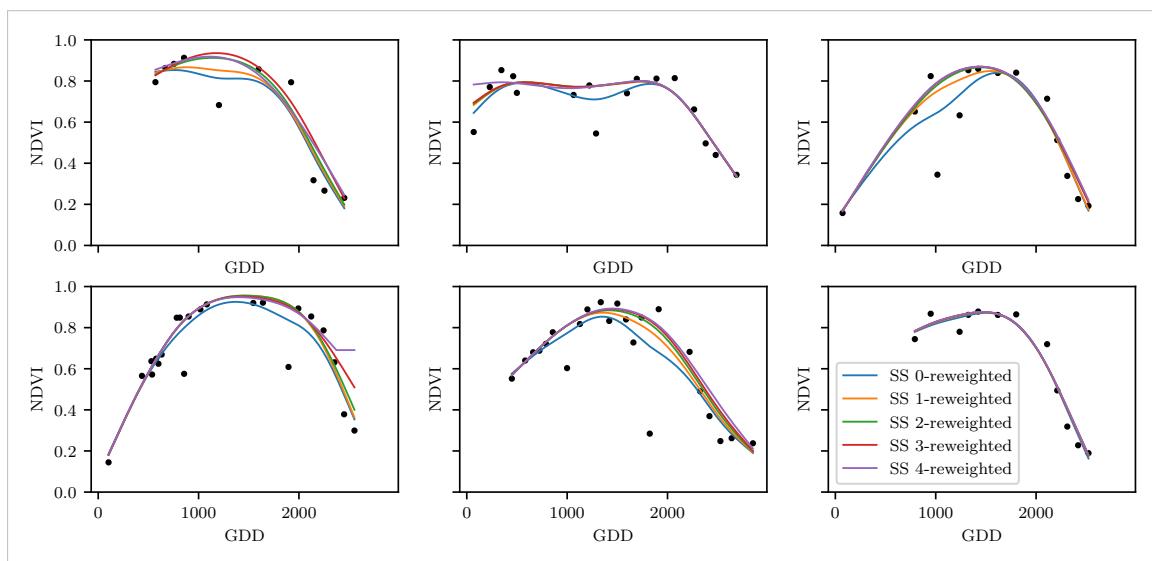


Figure 3.5: Smoothing splines fitted to different (SCL45) NDVI TS. Iterations of a robustifying refit (as indicated in section 3.5) are also displayed.

489 **3.5.3 Upper Envelope Approach - Penalty for Negative Residuals**

490 If we artificially increase the negative residuals in 3.5.0.1 by multiplying (e.g., factor 2),  
 491 the corresponding points will get less weight in the next iteration. This allows us to create  
 492 an interpolation that resembles an upper envelope. Intuitively, this upper envelope can be  
 493 thought of as a sheet that is laid on top of the points.

494 This approach is based on the premise that we tend to underestimate the NDVI (Cao  
 495 et al., 2018). Since we want to develop a general method that is in principle not related  
 496 to the NDVI, we will not pursue this approach further.

497 **3.6 Performance Assessment**

498 Next, we will benchmark the in section 6.1.2 preselected IMs with and without robustifi-  
499 cation. For this, we will use the same technique as we did for the parameter determination  
500 in section 3.4. On  $B_\lambda$  we apply the RMSE and different quantiles.

501 The results are presented in section 5.1 and are discussed in section 6.1. The double logistic  
502 turns out to be the best convincing parametric method and from the non-parametric  
503 methods we choose the SS.

504 **Chapter 4**

505 **NDVI Correction**

506 Let's remind ourselves that the data from the S2 satellites is distributed with an SCL and  
507 we therefore have some evidence about what is observed at each pixel for each sampled  
508 time (cf. table 2.2). So far, we have only considered points, labeled as cloud- and shadow-  
509 free (SCL45). However, we remind ourselves of the satellite images in figure 2.3d, where  
510 we had cloudy images despite the 'vegetation' label and see vegetation in figure 2.3e even  
511 though we are supposed to observe 'cirrus clouds'.

512 In this chapter, we will try to improve our NDVI interpolation by not relying only on the  
513 observed NDVI, but by training our own model to correct the NDVI using all S2 bands.  
514 For this, we introduce several statistical modelling approaches and discuss the strengths  
515 and weaknesses for each of them. After correcting the observed NDVI, we will assess the  
516 uncertainties of our corrections and translate them into weights. These will be used for  
517 the subsequent interpolation. This step-by-step procedure is illustrated by the figure B.4  
518 in the appendix. Finally, we will evaluate which combination of IMs and correction model  
519 performs the best.

520 **4.1 Considering other SCL Classes**

521 In figure 4.1 we plot the observed NDVI and notice that some blue points which correspond  
522 to the SCL-class 10 (thin cirrus clouds) follow the interpolated line closely. Hence, they  
523 might be useful in improving an interpolation fit.

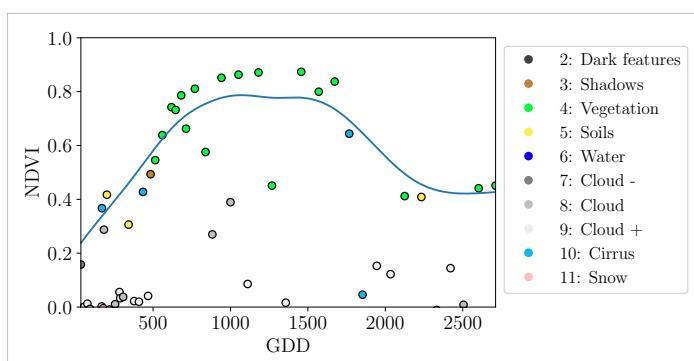


Figure 4.1: A smoothing splines fit considering green and yellow points (SCL45).

524 To get an impression of whether there is some useful information contained in non-SCL45

525 observations, we would like to compare the observed NDVI with the true NDVI. But since,  
 526 we do not have any ground truth data, we will make the following assumption:

527 **Assumption 4.1.0.1.** The “true” NDVI value at time  $t$  can be successfully estimated by  
 528 robustified LOOCV interpolation using high-quality observations. That is, the interpolated  
 529 value (using a robustified IM from chapter 3) considering the points  $P^{SCL45} \setminus P_t$ . In the  
 530 following, we will call this estimate the “true”-NDVI.

531 We would like to get an idea if there is any information that can be recovered from non-  
 532 SCL45 observations. For that, we will check for the other SCL-classes if there is a relation  
 533 between the “true” NDVI (derived with robustified SS) and the observed NDVI. Thus, we  
 534 pair each “true” NDVI with its observed one, collect all pairs, and create a scatter plot  
 535 for each SCL-class in fig 4.2. As expected, the “true” and the observed NDVI seem to be  
 536 highly correlated for SCL45. But we can also detect some patterns of correlation in the  
 537 SCL-classes 2, 3, 7, 8 and 10.

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

## 542 4.2 Correction Models

543 For training an NDVI correction model, we require ground-truth data which we will aim  
 544 to model using informative covariates. Since ground-truth NDVI data is not available,  
 545 we will again use the assumption 4.1.0.1 and use the “true” NDVI instead. There is no  
 546 canonical answer to the question of which covariates we should use. It is a tradeoff between  
 547 simplicity, generalizability and performance (with the danger of overfitting). Our desire  
 548 with the NDVI correction is to develop a product that is simple to use and understand.  
 549 Therefore, in the subsequent, we will only take the spectral data of the satellite (i.e., all  
 550 the bands) and the observed NDVI derived from it as covariates. We organize the chosen  
 551 covariates in the design matrix  $X^1$ , where each row corresponds to a  $P_t$  (i.e., a pixel at a  
 552 time  $t$ ) and each column to one covariate.

553 In the following, we will introduce different approaches, to model the relationship between  
 554 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  
 555 study the basic OLS. Second, we look at the LASSO, an penalized adaptation of the  
 556 OLS which is known to successfully deal with highly correlated covariates. Afterwards,  
 557 GAMs are introduced which model the response similar to OLS but allow for non-linear  
 558 relations. Last but not least, we discuss RF and MARS, which are both flexible modelling  
 559 approaches.

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

### 562 4.2.1 Ordinary Least Squares (OLS)

563 The OLS is a linear model that aims to minimize the sum of the squared residuals. We  
 564 assume a linear relationship between  $y$  and  $X$  and allow for Gaussian noise. That is:

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

---

<sup>1</sup>Strictly speaking, we include also the intercept and introduce one dummy variable for each SCL-class.

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

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

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

Advantages	Disadvantages
<ul style="list-style-type: none"> <li>— Simple method with good interpretability of coefficients.</li> <li>— Computationally cheap.</li> </ul>	<ul style="list-style-type: none"> <li>— Catches only linear relationships.</li> <li>— No integrated variable selection.<sup>2</sup></li> </ul>

568 **4.2.2 Least Absolute Shrinkage and Selection Operator (LASSO)**

569 The LASSO can be similarly expressed than the OLS but adds a penalty to the minimization  
570 problem:

$$\hat{\beta}_\lambda = \arg \min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 = \arg \min_{\beta \in \mathbb{R}^p \text{ and } \|\beta\|_1 < \lambda} \|y - X\beta\|_2^2. \quad (4.2.2.1)$$

571 Even though we do not have a closed form solution for equation (4.2.2.1) we can solve  
572 it easily via optimization, since the function  $\beta \in \{\beta \in \mathbb{R}^p | \|\beta\|_1 < \lambda\} \mapsto \|y - X\beta\|_2^2$  is  
573 continuous and convex.

574 Tibshirani (2011) shows that the LASSO solution tends to be sparse. That is  $\beta_i = 0$  for  
575 most  $i = 1, \dots, p$ . The larger  $\lambda$ , the more  $\beta_i = 0$  and hence the simpler the resulting  
576 model.

577 In order to know which  $\lambda$  to choose, we try a huge range of possible values. For each  
578  $\beta_\lambda$ , we calculate the cross-validated  $RMSE_\lambda$ <sup>4</sup> (and its standard deviation  $\sigma_\lambda$  using the  $k$   
579 folds) and define the  $\lambda$  with the smallest corresponding  $RMSE_\lambda$  as  $\lambda_{min}$ . From here we  
580 choose the largest  $\lambda$  for which the  $RMSE_\lambda$  is smaller than  $RMSE_{\lambda_{min}} + \sigma_\lambda$ . This yields  
581 a simpler model while keeping the  $RMSE$  reasonable model.

582 We will apply the LASSO using the selected covariates in section 4.2 and their second  
583 degree of interactions.<sup>5</sup>

Advantages	Disadvantages
<ul style="list-style-type: none"> <li>— Usually yields a sparse solution. This tends to give better generalizability (prediction performance on unseen data).</li> <li>— Successfully deals with correlation in covariates.</li> <li>— Interpretable results.</li> </ul>	<ul style="list-style-type: none"> <li>— Estimate is biased.</li> <li>— Computationally expensive.</li> </ul>

<sup>3</sup>The last two terms are equivalent by lagrangian optimization.

<sup>4</sup>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.

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

584 **4.2.3 General Additive Model (GAM)**

585 GAMs as described in [Hastie and Tibshirani \(1987\)](#) are a special case of Projection Pursuit  
 586 Regression, where only the  $p$  directions parallel to the coordinate axes are considered. The  
 587 result is different to a linear model since the coordinate functions are not restricted to be  
 588 linear but are assumed to be non-parametric functions. The model can be written as:

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

589 To estimate the non-parametric functions, we can use SS (ref sec. [3.3.6](#)). For this let  $\mathcal{S}_j$   
 590 be the function that takes some  $z \in \mathbb{R}^n$  and returns the SS fitted to  $(X_{:,j}, z)$  where the  
 591 smoothing parameter is optimized by LOOCV<sup>7</sup>. Since we cannot fit all  $g_j$  simultaneously,  
 592 we will use a strategy named Backfitting. We basically cycle through the indices  $1, \dots, p$   
 593 and refit  $\hat{g}_j$  each time. The following illustrates the procedure:

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

594 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.

595 **4.2.4 Random Forest (RF)**

596 To define a random Forest introduced by [Breiman \(2001\)](#) we will first define what a Tree  
 597 is. A (decision) Tree is a graph  $(V, E)$  without circles, a distinct root node, every node  
 598 has at most two children and every leaf has a value assigned to it. At each node there  
 599 is a boolean condition testing if one variable is greater than some value and a pointer to  
 600 one child depending on the boolean value. To evaluate a tree we start at the root node,  
 601 test the boolean expression and go to the node indicated by the resulting pointer. This  
 602 we repeat until we end up at a leaf-node, where we return the value assigned to it.

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

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

<sup>7</sup>For efficiency an proxy of the LOOCV is used called generalized cross validation.

<sup>8</sup>For computational reasons, we will only use splits along one covariate. So we ‘cut’ our covariate space into rectangles.

<sup>9</sup>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).

606 To build a Random Forest we will bootstrap-aggregate<sup>10</sup> many such Trees<sup>11</sup>. The prediction  
 607 of the Random Forest for a new point  $x$  is then the mean of the predictions from all  
 608 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.

#### 609 4.2.5 Multivariate Adaptive Regression Splines (MARS)

610 A MARS model as introduced in [Friedman \(1991\)](#) can be described by

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

611 where the  $h_m$  are simple functions (explained later) and the  $\beta_m$  are estimated via Least  
 612 Squares.

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

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

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

$$\mathcal{C} := \{(h(\cdot)b_1(\cdot), h(\cdot)b_2(\cdot)) \mid h \in \mathcal{M}, (b_1, b_2) \in \mathcal{B}\} \quad (4.2.5.1)$$

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

621 Finally, to avoid overfitting, we prune the set  $\mathcal{M}$  by optimizing a LOOCV score.<sup>12</sup>

622 To reduce computational complexity, we follow the recommendation from [Stephen \(2021\)](#)  
 623 and restrict  $h$  in equation (4.2.5.1) to be of degree one (so it is also in a pair of  $\mathcal{B}$ ).  
 624 Consequently,  $\mathcal{C}$  contains functions with a degree of at most 2.

<sup>10</sup>That is we will sample (with replacement) several times  $n$  observations from our original data and fit a Tree to each such sample.

<sup>11</sup>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.

<sup>12</sup>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  $\mathcal{M} \setminus \{h\}$ . We discard the function that – when excluding from  $\mathcal{M}$  – leads to the best LOOCV score.

Advantages	Disadvantages
<ul style="list-style-type: none"> <li>— Catches non-linear relationships.</li> <li>— Interpretability via functions in <math>\mathcal{M}</math> and their coefficients.</li> <li>— Allows for interactions with variable selection.</li> </ul>	<ul style="list-style-type: none"> <li>— Computationally expensive (can be reduced by restricting the degree of interactions).</li> </ul>

### 625 4.3 Weighted Interpolation

626 Once we corrected the NDVI using the models described in the previous section, we are left  
 627 with the problem that not every correction is equally reliable.<sup>13</sup>. Hence, we are interested  
 628 in a measure of how uncertain an estimate is. We achieve this analogously as we corrected  
 629 the NDVI, by replacing the response (NDVI-“true”) with the absolute residuals  $v := |y - \hat{y}|$   
 630 and modeling their relationship with the covariates defined by  $X$ . In this way, we obtain  
 631 a model for the absolute residuals  $v$  and the estimator  $\hat{v}$ .

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

$$w_{\tau}^{(P)} := \frac{1}{R} \frac{1}{\hat{v}_{\tau}^{(P)}}, \quad \text{for } \tau = 1, \dots, n_P \quad (4.3.0.1)$$

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

### 639 4.4 Resulting Interpolation Strategies (ISs)

640 We have developed the following procedure to obtain a new interpolation (keyword-wise):  
 641 i.) LOOCV Interpolation (+ robustify?) to get “true” NDVI  
 642 ii.) Correction  
 643 iii.) Uncertainty estimation  
 644 iv.) Interpolation (+ robustify?)

645 At each step we have a choice, more precisely:

- Interpolation: Smoothing Splines / Double Logistic
- Robustify: Yes / No
- Correction & uncertainty estimation: RF / OLS – considering only SCL-classes / OLS – considering all selected covariates / MARS / GAM / LASSO / no correction.

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

<sup>13</sup>One correction is illustrated in the figure B.4f. In this figure, the outer points (labeled as clouds) have a large scatter.

- 652 — We use the same IM each time.  
 653 — Either we robustify both times, or we do not robustify at all.  
 654 — We use the same underlying method for correction and uncertainty estimation.  
 655 In this fashion, we obtain 28 distinct ISs, which we will benchmark in the next section.

## 656 4.5 Evaluation via (relative) Yield Prediction Error (relative YPE)

657 In this section, we introduce the relative YPE and utilize it to evaluate the 28 ISs from  
 658 section 4.4. The fundamental assumption is that the closer the interpolated NDVI TS is  
 659 to the true one, the better it can be used to determine crop yield. Implicitly, we believe  
 660 that an NDVI TS that better models yield will incorporate more true information about  
 661 the underlying vegetation. Therefore, we want to determine a comparable YPE for each  
 662 IS and choose it as a benchmark criterion. This is an objective measure, since we have  
 663 not considered crop yield in any of our previous steps. Moreover, this criterion is justified  
 664 by the fact that yield estimation has been a motivation for the interpolation.

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

$$YPE = \frac{\sqrt{\sum_{i=1}^n (y_i - \hat{y}_i)^2}}{\bar{y}},$$

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

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

- Maximum slope
- Minimum slope
- Integral<sup>15</sup> over all
- Peak (i.e., maximal NDVI)
- GDD for the Peak
- Integral<sup>15</sup> up to the peak
- Integral<sup>15</sup> after peak
- Integral<sup>15</sup> from 0-685 GDD
- Integral<sup>15</sup> from 685-1075 GDD

673 For the choice we were inspired by (cf. table 2 in Kamir et al. (2020)). However, we  
 674 deliberately omit any statistic that involves the minimum (e.g., the NDVI-range), since  
 675 we regard the minimum as a very error-prone measure due to the large influence of clouds  
 676 in the TS.

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

---

<sup>14</sup>We will use the matrixes derived in section 4.5.

<sup>15</sup>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 2.3a and 2.3f with their NDVI in plot 2.3g).

680 the integrated OOB estimates<sup>16</sup>  $\hat{y}$ . Note that the choice of the modeling approach does not  
681 matter much, as long as it is general enough (i.e., able to approximate any function) and  
682 we use the same one for each IS. Finally, for each IS, we calculate the YPE and describe  
683 the results in section 5.2.

---

<sup>16</sup>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  $\frac{1}{e}$ ).

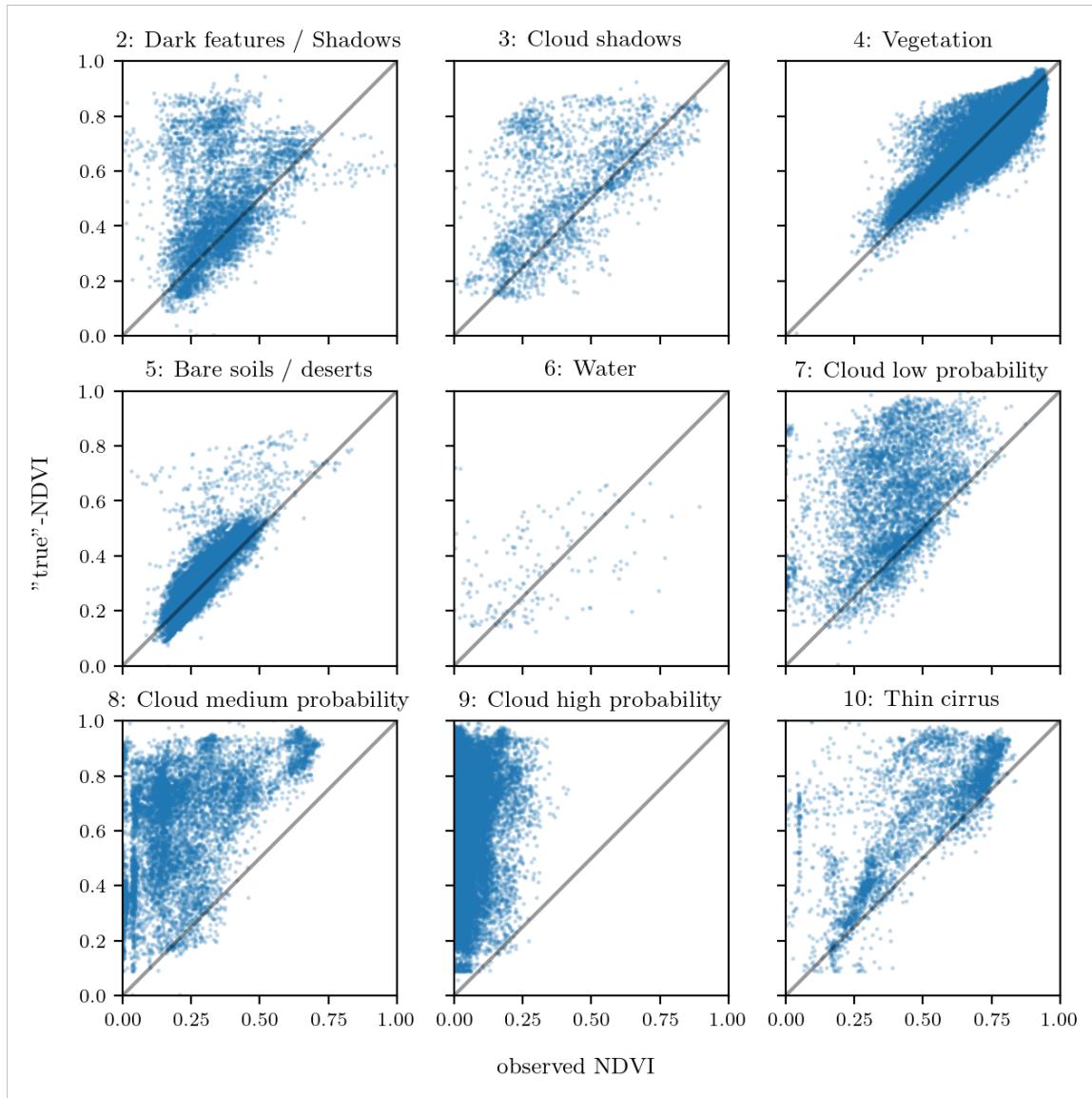


Figure 4.2: For each SCL class, we compare the true NDVI with the observed NDVI. (The true NDVI was estimated with LOOCV smoothing splines, and we used all observations of 10% of the total pixels.)

684 **Chapter 5**

685 **Results**

686 **5.1 Goodness of Fit for Selected IMs**

687 Table 5.1 benchmarks the selected<sup>1</sup> IMs (on  $P^{SCL45}$ ) with respect to various score func-  
 688 tions. The score functions take the absolute values of the LOOCV residuals and summarize  
 689 them in a number (the smaller, the better). For each of the 5 selected IMs, we consider  
 690 the basic and the robustified (see section 3.5) version.

Table 5.1: Comparing the goodness of fit for selected IMs (on  $P^{SCL45}$ ) measured with the score functions (that take the LOOCV residuals as input) listed in the left column.  $q_X$  denotes here the  $X\%$  quantile. Colored rowwise.

	SS	LOESS	DL	BS	FS	$SS^{\text{rob}}$	$\text{LOESS}^{\text{rob}}$	$DL^{\text{rob}}$	$BS^{\text{rob}}$	$FS^{\text{rob}}$
RMSE	0.063	0.061	0.061	0.074	0.075	0.070	0.065	0.065	0.079	0.208
qtile50	0.036	0.034	0.027	0.043	0.031	0.032	0.031	0.022	0.037	0.049
qtile75	0.063	0.061	0.051	0.077	0.058	0.061	0.057	0.044	0.070	0.099
qtile85	0.080	0.079	0.070	0.098	0.083	0.081	0.076	0.063	0.094	0.158
qtile90	0.092	0.092	0.088	0.112	0.108	0.097	0.090	0.082	0.113	0.226
qtile95	0.119	0.115	0.122	0.142	0.161	0.132	0.115	0.124	0.157	0.375

691 DL is the best among both robustified and non-robustified with respect to most of the  
 692 score functions used (all except q95) and is especially superior to the other parametric  
 693 approach, which is FS. Especially the robust FS performs poorly. The LOESS dominates  
 694 (i.e., is superior on every score function) all other non-parametric methods, but is closely  
 695 followed by the SS. The BSPL, on the other hand, is the worst non-parametric method  
 696 tested here.

697 **5.2 YPE for Tested ISs**

698 The YPE for the 28 (in section 4.4) chosen ISs is given in table 5.2. We note that  
 699 robustification, does not improve the quality of the fit (measured via the YPE) in most  
 700 cases. In addition, SS (rob) tend to better than DL(rob) in terms of YPE, especially if no  
 701 correction is made. The IS that leads to the lowest YPE is the OLS<sup>SCL</sup> with SS. Given  
 702 that the OLS<sup>SCL</sup> models have very good interpretability, we also present the regression

---

<sup>1</sup> For the discussion which methods have been selected cf. section 6.1.2.

Table 5.2: Relative YPE for various ISs. For the non-relative YPE and the coefficient of determination ( $R^2$ ) cf. table B.1 and B.2.

	RF	OLS <sup>SCL</sup>	OLS <sup>all</sup>	MARS	GAM	LASSO	no corrections
SS	0.155	0.140	0.143	0.142	0.142	0.142	0.149
SS <sup>rob</sup>	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 <sup>rob</sup>	0.157	0.153	0.152	0.145	0.148	0.150	0.157

703 equations below. The corrected NDVI is calculated using

$$\begin{aligned} \text{NDVI}_{\text{corr}} = & 0.711 \text{NDVI}_{\text{observed}} + \mathbb{1}_{SCL=2} 0.215 + \mathbb{1}_{SCL=3} 0.237 + \mathbb{1}_{SCL=4} 0.210 \\ & + \mathbb{1}_{SCL=5} 0.116 + \mathbb{1}_{SCL=6} 0.162 + \mathbb{1}_{SCL=7} 0.327 + \mathbb{1}_{SCL=8} 0.474 \quad (5.2.0.1) \\ & + \mathbb{1}_{SCL=9} 0.575 + \mathbb{1}_{SCL=10} 0.306 + \mathbb{1}_{SCL=11} 0.512 \end{aligned}$$

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

$$\begin{aligned} \widehat{\text{abs}}(\text{NDVI}^{\text{"true"}} - \text{NDVI}_{\text{corr}}) = & -0.133 \text{NDVI}_{\text{observed}} + \mathbb{1}_{SCL=2} 0.186 + \mathbb{1}_{SCL=3} 0.185 \\ & + \mathbb{1}_{SCL=4} 0.146 + \mathbb{1}_{SCL=5} 0.089 + \mathbb{1}_{SCL=6} 0.167 \\ & + \mathbb{1}_{SCL=7} 0.203 + \mathbb{1}_{SCL=8} 0.181 + \mathbb{1}_{SCL=9} 0.173 \\ & + \mathbb{1}_{SCL=10} 0.180 + \mathbb{1}_{SCL=11} 0.172 \end{aligned} \quad (5.2.0.2)$$

706 In the equation 5.2.0.1, we notice the strongest upwards correction for SCL classes 8, 9 and  
707 11 (correspond to ‘medium probability clouds’, ‘high probability clouds’ and ‘thin cirrus  
708 clouds’). The estimated absolute residuals, however, are the smallest for SCL classes 4 and  
709 5 (‘vegetation’ and ‘bare soil’). Furthermore, the higher the observed NDVI the lower are  
710 the estimated absolute residuals.

711 For the R-output of the `summary` function of the two models, we refer to the appendix B.3.1.

---

<sup>2</sup>  $\mathbb{1}$  is also called an indicator function or characteristic function in mathematics.

712 **Chapter 6**

713 **Discussion**

714 This chapter is a collection of arguments and justifications regarding various decisions.  
715 First, we examine IMs for compatibility with data gaps and argue choices for selected  
716 IMs. Second, we identify the best IS and investigate issues that have arisen in the context  
717 of the NDVI correction.

718 **6.1 IMs**

719 **6.1.1 Data Gaps in Time Series**

720 NW estimates the value for  $t$  by relating to the points near  $t$ . To determine what “near”  
721 means, a bandwidth  $h$  is used (cf. equation 3.3.1.2). This gets problematic as soon as the  
722 data gaps become larger than  $h$ , since in this case no points are left that are considered  
723 to be close to  $t$ .

724 Regarding the GK, we expect that due to the stationarity assumption, the interpolation  
725 will always tend to the mean if data gaps are present (cf. figure 3.3).

726 Since the SG requires equidistant points, it follows that data gaps will break it. The  
727 linear interpolation, that is supposed to recover this, we consider as not being a satisfying  
728 solution.

729 We do not trust the FS interpolation if there are noticeable data gaps. On the one hand,  
730 it corresponds to our experience that the curve can escape strongly there (cf. figure 3.1).  
731 On the other hand, the unreliability is illustrated by the poor values in table 5.1 for  
732 the robustified variant. These are meaningful in describing the ability to cope with data  
733 gaps, since more data points are ignored during the robustification and thus data gaps are  
734 simulated.

735 Similarly, for SS, LOESS, DL and BS we compare the values in table 5.1 between the  
736 robustified and non-robust variant. We find that the robust variant does not differ strongly  
737 from the non-robust variant (unlike as for FS). Thus, we conclude that these methods do  
738 not have systematic failures.

739 Regarding the LOESS, we observe in the figure B.1 in plot (c) a strange peak between  
740 the first and second observation. This peak is due to the local weighting. In case of data  
741 gaps, the weights can attain non-intuitive values. For instance, the first data point in the

742 plot, although adjacent to the peak, is given a low weight compared to the points to the  
 743 right of the peak (for estimating the value at this peak).

744 In our experience, the DL handles data gaps well, but it may happen that the model  
 745 describes the NDVI increase as abrupt. This however was fixed, by bounding the first  
 746 derivative (cf. section 3.2.3).

### 747 6.1.2 Preselection

748 We shall now justify our preselection of the IMs tested in section 3.6. We decided against  
 749 NW Because of its systematic errors at peaks and valleys. Moreover, this method handles  
 750 data gaps poorly (cf. 6.1.1). Moreover, UK will not be considered since the underlying  
 751 assumptions are not met and therefore a systematic bias is introduced. On top of that,  
 752 maximum likelihood parameter estimation occasionally fails. Also, we do not include the  
 753 SG in the next selection, since we see it as a special case of LOESS. The remaining IMs  
 754 are thus SS LOESS DL BS and FS.

### 755 6.1.3 Candidate Selection

756 Given that DL convinces regarding most of the selected score functions in table 5.1 we will  
 757 certainly investigate this method in chapter 4. Moreover, we see that the robustification  
 758 mostly improved the score regarding the 50, 75, 85, and 90 % Quantiles. Only for the  
 759 outlier-sensitive score functions (RMSE and q95)<sup>1</sup> we notice significant worsening (we  
 760 consider the robust FS separately in section 6.1.1). Consequently, we will also use the  
 761 robustification in section 4. In order to not only rely on the form assumptions of the  
 762 DL, we further choose a non-parametric method for further consideration. Despite the  
 763 LOESS slightly dominating the SS in table 5.1, we choose the SS. This is due to the  
 764 strange behavior of the LOESS in case of data gaps (see section 6.1.1) and the good  
 765 interpretability of the SS using the minimization function 3.3.6.1.

## 766 6.2 NDVI Correction

### 767 6.2.1 Choose IS

768 The evaluation of various ISs via the YPE (cf. section 5.2) clearly shows that SS are  
 769 better suited than DL for yield estimation. Moreover, it seems surprising that robustifi-  
 770 cation tends to worsen the results, despite reducing LOOCV residuals in most cases (cf.  
 771 section 5.1). We suspect that the correction models handles outliers by themselves (by  
 772 correcting or down-weighting them) and thus do not benefit from an external robustifica-  
 773 tion. Indeed, for OLS<sup>SCL</sup> we see in equation 5.2.0.2 that the smaller the observed NDVI  
 774 of a point, the larger the estimated residual — yielding a lower weight. This is consistent  
 775 with our experience that outliers usually suggest a too small NDVI value. Our suspicion  
 776 is consistent with the fact that if we do not correct, robustification produces a marginal  
 777 improvement. By using the best IS with correction (SS+OLS<sup>SCL</sup>), rather than the best  
 778 IS without correction (SS<sup>rob</sup>), we gain  $(0.148 - 0.140)/0.148 = 5.4\%$  of information about  
 779 the underlying vegetation — where 100% would allow us to model the yield perfectly.

780 Note that the results discussed here depend strongly on the link function used. Once we  
 781 change this, we should also repeat this analysis.

---

<sup>1</sup>For the RMSE one outlier is enough to take away the usefulness of the statics, in the case of q95 it is enough if 5% of the data are corrupt to break the statics.

### 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 Perich et al. (2022). There, a non-relative YPE 1.00 [t/ha] was obtained using weather data in addition to NDVI TS. Since our error is only about 3.3% larger (cf. table B.1), we consider our results to be competitive. Especially as we did not use meteorological data aside from the time-scale transformation (cf. section 2.4). 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.) Uncertainty in satellite images through “measurement errors” introduced via clouds and other atmospheric effects.
- iv.) Uncertainty introduced by assumed homogeneity of vegetation on pixel resolution.
- v.) Uncertainty introduced by interpolating NDVI TS (especially when long data-gaps are present).

Furthermore, even if we would have a perfect NDVI curve, it contains only a fraction of the information about the underlying vegetation. Nonetheless, Perich et al. (2022) manages to explain up to 86% of the variance in crop yield with only the NDVI TS and weather 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 neighbouring pixels from the test data and consequently exhibit high correlations (in yield and NDVI). We suspect that overfitting via high-correlation pixels is responsible for these high values. On the other hand, the authors observe poor results for cross-year-validation<sup>2</sup> (table 6) and 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<sup>3</sup>. Nevertheless, we claim, that our results are not corrupted by the correlation of neighboring pixels. This is because our result is not a ‘good’ YPE, but the selected IS. So all tested ISs 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 4.1.0.1 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.

<sup>2</sup>By cross-year-validation we understand a cross validation with respect to the RMSE, where each year represents a single fold.

<sup>3</sup>By cross-field-validation we understand a cross validation with respect to the RMSE with a partitioning  $\mathcal{F} = \{F_1, \dots, F_m\}$  such that all pairs of pixels from the same year with the same field ID, it holds that both pixels are in the same  $F_k$ .

820 **6.2.4 Using Additional Covariates**

821 In section 4.2 we have only used covariates derived from spectral data. Even though we  
822 included meterological data from Perich et al. (2022) in our implementation, we decided  
823 against using this data. On the one hand we have the problem that we have practically too  
824 few observations (we observe only 5 years). Moreover, we expect the weather in our study  
825 region to be rather homogeneous which is suggested by the fact that the weather data  
826 published by Meteoswiss are for a grid with a resolution of 1 km. On the other hand, we  
827 want the underlying model not to learn improper relationships. For example, the model  
828 might automatically predict a high NDVI for a day in summer (detected by high GDD  
829 / many sunshine hours / high temperature) just because it is “used” to observing a lot  
830 of vegetation in summer. Including temporally (e.g.,  $P_{t-1}$  and  $P_{t+1}$ ) and geographically  
831 adjacent pixels would likely improve performance. However, for simplicity, we omit it  
832 here<sup>4</sup>.

---

<sup>4</sup>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 is a border-pixel with some adjacent pixels outside of the field.

833 **Chapter 7**

834 **Conclusion**

835 In this thesis, we investigated how to model vegetation dynamics through NDVI TS derived  
836 from satellite images. The Scene Classification Layer (SCL), supplied by the European  
837 Space Agency, played a key role in this process. The major challenges faced were how to  
838 deal with contaminated observations (due to clouds or shadows) and how to interpolate  
839 the observed NDVI values. A summary of the IMs considered can be found in the table 3.

840 To make the IMs more robust to contaminated observations (outliers) that remained af-  
841 ter SCL filtration, we generalized an iterative technique. After an initial fit, in each  
842 iteration we give less weight to observations with comparatively large residuals and then  
843 perform a weighted interpolation (see section 3.5). However, after too many iterations,  
844 non-contaminated points might get ignored (i.e., given a zero weight). The greatest im-  
845 provements, on the other hand, were perceived after the first iteration (see figure 3.5).

846 Filtering the observations contaminated by clouds and shadows via SCL introduces data  
847 gaps, especially in winter. Therefore, we aim for IMs that handle such data gaps well.  
848 The Nadaraya-Watson kernel estimator struggles when there are no or too few points  
849 in the window of interest; Universal Kriging is biased towards the mean, particularly in  
850 environments with no data (cf. figure 3.3); 2cd order Fourier series can deviate strongly  
851 within data gaps (cf. figure 3.1) and the Savitzky-Golay filter depends on equidistant  
852 observations (cf. section 6.1.1). Occasionally, a generalization of the Savitzky-Golay filter  
853 — the Locally Weighted Regression — has also shown surprising behavior in data gaps  
854 (cf. figure B.1).

855 In contrast, the latter performed well in Leave-One-Out-Cross-Validation (LOOCV) (cf.  
856 table 5.1). Nevertheless, we prefer the Smoothing Splines (SS) as they perform only slightly  
857 worse there, but produce a much smoother curve (cf. figure 3.5 and B.1). SS flexibly  
858 approximate the data while keeping curvature low (cf. equation 3.3.6.1). B-splines, on  
859 the other hand, were worse than SS with respect to every score function tested, and their  
860 smoothing mechanism is also less interpretable. However, the best performing method  
861 here is the approximation by a Double logistic (DL), which makes strong assumptions  
862 about the shape of the NDVI curve. Problems for the parameter estimation of the DL  
863 (and the Fourier series) have been resolved by restricting the parameter space by generous  
864 but realistic values. Problems with overfitting in universal kriging were overcome by  
865 determining the variogram parameters for a subsample of NDVI TS and finally using the  
866 median of each parameter. In the end, we choose DL and SS as our preferred IMs.

867 Question: more details for the justification of the interpolation candidates?

868 The traditional answer to the question of how to deal with contaminated observations is  
869 that we only consider observations that are labeled as vegetation or bare soil by the SCL  
870 (SCL45). The unreliability of this labeling, however, is illustrated in figure 2.3. Moreover,  
871 filtered observations (non-SCL45) might still contain valuable information (see section 4.1).  
872 Therefore, we do not adhere to traditional (SCL) filtration but instead consider all ob-  
873 servations and correct the observed NDVI with uncertainty estimation. For this, we use  
874 statistical models that take additional information such as the remaining spectral bands,  
875 the current SCL label and the observed NDVI into account. But before we interpolate the  
876 corrected NDVI values, we assign a weight to each observation, corresponding to its un-  
877 certainty. The uncertainty is estimated analogously as the NDVI has been corrected. By  
878 combining different IMs (with and without robustification) with various statistical models,  
879 we obtain 28 different Interpolation Strategies (ISs) (see section 4.4). To assess which of  
880 these ISs is best, we assume that the better the IS, the better it allows interpolated NDVI  
881 TS to predict yield. Surprisingly, the best strategy is the one with non-robust SS and the  
882 simplest static model considered, which uses only the observed NDVI and SCL classifica-  
883 tion. Let us recapitulate the best IS: First, we estimate the “true” NDVI (c.f. assumption  
884 4.1.0.1) using SS via LOOCV. Then obtain the corrected NDVI using the OLS<sup>SCL</sup> model  
885 (cf. equation 5.2.0.1). Subsequently, we estimate the absolute error with the OLS<sup>SCL</sup>  
886 model (cf. equation 5.2.0.1) and thereby obtain weights which are supposed to reflect the  
887 reliability of the corrected NDVI (cf. equation 4.3.0.1). Finally, we perform a weighted  
888 interpolation with SS.

889 For evaluating the generalized robustification technique, we used raw LOOCV performance  
890 on the one hand, and the ability to model plant growth for crop yield estimation on the  
891 other hand. While the robustification is not part of the best IS, it narrowly misses this  
892 target. In contrast, we see in table 5.1 that robustification leads to smaller LOOCV  
893 residuals in most cases. That is (with the exception of the Fourier approximation) the  
894 50% and 75% quantiles of the absolute residuals are smaller for the robustified ones. Hence,  
895 when we expect contaminated observations, we advise to robustify the interpolation.

896 As to the question which IM we recommend, we consider two cases. If one only intends  
897 to fit a curve to the data as precisely as possible, we recommend the robustified DL, since  
898 it minimizes the LOOCV residuals in most cases (cf. table 5.1). In the event that one  
899 requires an interpolation that contains as much information about the plant as possible,  
900 we recommend the SS. This recommendation is especially valid if we traditionally consider  
901 only SCL45 observations without correcting the proposed NDVI. However, we recommend  
902 the abovementioned IS with NDVI correction, because otherwise 5.4% of the information  
903 about the vegetation will be lost from the NDVI TS (cf. section 6.2.1). In light of all the  
904 sources of error (cf. section 6.2.2) and the fact that we only consider the NDVI TS, we  
905 consider the 5.4% to be a solid improvement.<sup>1</sup>

<sup>1</sup>The 5.4% corresponds to the reduction in variance in the crop yield estimate with the corrective IS compared to a traditional SS interpolation. 100% would thus suggest that we could perfectly predict yield from the interpolated NDVI curve (despite all the sources of error mentioned above).

## 906 7.1 Future Work

## 907 7.1.1 Time Series Correction-Interpolation as a General Method

908 Throughout this thesis, we developed a correction and IM for the NDVI. However, we never  
909 used features of the NDVI. Only the parameter estimated via cross-validation in chapter 3.4  
910 depends on the scale of the TS. For simplicity, we could thus determine the parameter using  
911 Generalized Cross Validation (as Ripley and Maechler suggest). Therefore, our approach  
912 of interpolation and correction of TS can be applied to arbitrary TS as long as additional  
913 information is available. However, further research is required, to demonstrate the general  
914 usefulness of this approach.

## 915 Example: Cloud Correction with Uncertainty Estimation and Interpolation

916 This generalization can be used in particular for cloud correction. In the same manner as  
917 we corrected the NDVI TS in chapter 4, we can correct each spectral band and reunite  
918 the corrected bands with the uncertainties. If desired, the TS can also be interpolated  
919 before merging as in chapter 4.3. The resulting question would be how well this approach  
920 performs.

## 921 7.1.2 Minor Improvements

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

- 924 — **Data:** Method how combine harvester point data has been extrapolated to the grid  
925 could possibly be improved.
- 926 — **Data:** For computational reasons, we mostly considered all years and split the data  
927 (on the pixel level) randomly into a train/test set. A leave one year out cross  
928 validation might yield more accurate results.
- 929 — **Data:** We have not included the spectral bands that have a resolution of 60 m. But  
930 precisely these seem to be promising for cloud correction, since they are a proxy of  
931 the water (content and form) in the atmosphere.
- 932 — **Data:** Raiyani et al. (2021) presents an Machine Learing approach that supposedly  
933 improves the SCL and thus could improve our results that are based on the SCL.
- 934 — **NDVI Correction:** Explore the effect of different link and normalizing functions in  
935 section 4.3. Currently we run into the danger of some outer points getting nearly  
936 ignored just because one estimated absolute residual for some interior point is close  
937 to zero.
- 938 — **NDVI Correction:** Yield is not the only target variable of interest. Other variables  
939 like protein content could also be used in section 4.5 for the method evaluation.

940 

# Bibliography

- 941 (2007). Gaussian models for geostatistical data. In P. J. Diggle and P. J. Ribeiro (Eds.),  
942 *Model-Based Geostatistics*, pp. 46–78. New York, NY: Springer.
- 943 Atzberger, C. and P. H. Eilers (2011, September). A time series for monitoring vegetation  
944 activity and phenology at 10-daily time steps covering large parts of South America.  
945 *International Journal of Digital Earth* 4(5), 365–386.
- 946 Bailey, S. J. (2018, July). Using Growing Degree Days to Predict Plant Stages. pp. 8.
- 947 Beck, P. S. A., C. Atzberger, K. A. Høgda, B. Johansen, and A. K. Skidmore (2006,  
948 February). Improved monitoring of vegetation dynamics at very high latitudes: A new  
949 method using MODIS NDVI. *Remote Sensing of Environment* 100(3), 321–334.
- 950 Breiman, L. (2001, October). Random Forests. *Machine Learning* 45(1), 5–32.
- 951 Brockmann, M., T. Gasser, and E. Herrmann (1993, December). Locally Adaptive Band-  
952 width Choice for Kernel Regression Estimators. *Journal of the American Statistical  
953 Association* 88(424), 1302–1309.
- 954 Bühlman, P. and M. Maechler (2020, October). Computational Statistics.
- 955 Cai, Z., P. Jönsson, H. Jin, and L. Eklundh (2017, December). Performance of Smoothing  
956 Methods for Reconstructing NDVI Time-Series and Estimating Vegetation Phenology  
957 from MODIS Data. *Remote Sensing* 9(12), 1271.
- 958 Cao, R., Y. Chen, M. Shen, J. Chen, J. Zhou, C. Wang, and W. Yang (2018, November). A simple method to improve the quality of NDVI time-series data by integrating  
959 spatiotemporal information with the Savitzky-Golay filter. *Remote Sensing of Environ-  
960 ment* 217, 244–257.
- 962 Chandola, V. and R. R. Vatsavai (2010). Scalable time series change detection for biomass  
963 monitoring using Gaussian Processes. *Conference on Intelligent Data Understanding*,  
964 14.
- 965 Chen, J., P. Jönsson, M. Tamura, Z. Gu, B. Matsushita, and L. Eklundh (2004, June). A  
966 simple method for reconstructing a high-quality NDVI time-series data set based on the  
967 Savitzky–Golay filter. *Remote Sensing of Environment* 91(3), 332–344.
- 968 Cleveland, W. S. (1979, December). Robust Locally Weighted Regression and Smoothing  
969 Scatterplots. *Journal of the American Statistical Association* 74(368), 829–836.
- 970 Courault, D., L. Hossard, V. Demarez, H. Dechatre, K. Irfan, N. Baghdadi, F. Flamain,  
971 and F. Ruget (2021, July). STICS crop model and Sentinel-2 images for monitoring rice  
972 growth and yield in the Camargue region. *Agronomy for Sustainable Development* 41(4),  
973 49.

- 974 Eilers, P. H. C. (2003, July). A Perfect Smoother. *Analytical Chemistry* 75(14), 3631–3636.
- 975 ESA (2022a, August). Level-2A Algorithm Overview.  
976 <https://sentinels.copernicus.eu/web/sentinel/technical-guides/sentinel-2-msi/level-2a/algorithms>.
- 978 ESA (2022b, August). Sentinel-2. <https://sentinel.esa.int/web/sentinel/missions/sentinel-2>.
- 980 Friedman, J. H. (1991, March). Multivariate Adaptive Regression Splines. *The Annals of Statistics* 19(1), 1–67.
- 982 Gurung, R. B., F. J. Breidt, A. Dutin, and S. M. Ogle (2009, October). Predicting Enhanced Vegetation Index (EVI) curves for ecosystem modeling applications. *Remote Sensing of Environment* 113(10), 2186–2193.
- 985 Hastie, T. and R. Tibshirani (1987, June). Generalized Additive Models: Some Applications. *Journal of the American Statistical Association* 82(398), 371–386.
- 987 Henits, L., Á. Szerletics, D. Szokol, G. Szlovák, E. Gojdár, and A. Zlinszky (2022, January). Sentinel-2 Enables Nationwide Monitoring of Single Area Payment Scheme and Greening Agricultural Subsidies in Hungary. *Remote Sensing* 14(16), 3917.
- 990 Jaramaz, D., V. Perović, S. Belanovic Simic, E. Saljnikov, D. Cakmak, V. Mrvić, and L. Zivotic (2013, May). The ESA Sentinel-2 mission Vegetation variables for Remote sensing of Plant monitoring.
- 993 Kamir, E., F. Waldner, and Z. Hochman (2020, February). Estimating wheat yields in Australia using climate records, satellite image time series and machine learning methods. *ISPRS Journal of Photogrammetry and Remote Sensing* 160, 124–135.
- 996 Lyche, T. and K. Mørken (2005, January). Spline Methods.
- 997 McMaster, G. S. and W. W. Wilhelm (1997, December). Growing degree-days: One equation, two interpretations. *Agricultural and Forest Meteorology* 87(4), 291–300.
- 999 Omori, K., T. Sakai, J. Miyamoto, A. Itou, A. N. Oo, and A. Hirano (2021, April). Assessment of paddy fields' damage caused by Cyclone Nargis using MODIS time-series images (2004–2013). *Paddy and Water Environment* 19(2), 271–281.
- 1002 Perich, G., H. Aasen, J. Verrelst, F. Argento, A. Walter, and F. Liebisch (2021, January). Crop Nitrogen Retrieval Methods for Simulated Sentinel-2 Data Using In-Field Spectrometer Data. *Remote Sensing* 13(12), 2404.
- 1005 Perich, G., M. O. Turkoglu, L. V. Graf, J. D. Wegner, H. Aasen, A. Walter, and F. Liebisch (2022, July). Pixel-based crop yield mapping and prediction using spectral indices and neural networks on Sentinel-2 time series data.
- 1008 Raiyani, K., T. Gonçalves, L. Rato, P. Salgueiro, and J. R. Marques da Silva (2021, January). Sentinel-2 Image Scene Classification: A Comparison between Sen2Cor and a Machine Learning Approach. *Remote Sensing* 13(2), 300.
- 1011 Ripley, B. D. and M. Maechler. R: Fit a Smoothing Spline. <https://stat.ethz.ch/R-manual/R-patched/library/stats/html/smooth.spline.html>.
- 1013 Rouse, J. W. (1974, May). Monitoring the vernal advancement and retrogradation (green wave effect) of natural vegetation. Technical Report NASA-CR-139243.
- 1014

- 1015 Savitzky, A. and M. J. E. Golay (1964, July). Smoothing and Differentiation of Data by  
1016 Simplified Least Squares Procedures. *Analytical Chemistry* 36(8), 1627–1639.
- 1017 Schafer, R. W. (2011, July). What Is a Savitzky-Golay Filter? [Lecture Notes]. *IEEE*  
1018 *Signal Processing Magazine* 28(4), 111–117.
- 1019 Stephen, M. (2021, July). Earth: Multivariate Adaptive Regression Splines.
- 1020 Stöckli, R. and P. L. Vidale (2004, September). European plant phenology and climate  
1021 as seen in a 20-year AVHRR land-surface parameter dataset. *International Journal of*  
1022 *Remote Sensing* 25(17), 3303–3330.
- 1023 Strbac, O., M. Milanovic, and V. Ogrizovic (2017, July). Estimation the evapotrasnpira-  
1024 tion of urban parks with field based and remotely sensed datasets. pp. 13.
- 1025 Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: A retrospective.  
1026 *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73(3), 273–  
1027 282.

1028 **Appendix A**

1029 **Reproducibility**

1030 **A.1 Reproduce Results**

1031 For reproducibility of the whole computations, we refer to our codebase at:

1032 <https://github.com/LGraz/MasterThesis-Code>

1033 In order to reproduce our computations and results, set up the directory as described in the  
1034 README. The the ‘YieldMapping’ Data used, is published alongside [Perich et al. \(2022\)](#).

1035 Execute the computations via the script `./shell_scripts/reproduce.sh` and do not execute  
1036 the python and R files by hand (unless you follow the order in `./shell_scripts/reproduce.sh`).

1037 **A.2 R-Package**

1038 We also provide an R package for a general time series correction and interpolation if  
1039 additional data is available at:

1040 <https://github.com/LGraz/CorrectTimeSeries>

1041 In our case we consider the NDVI time series and the additional data consists of the unused  
1042 spectral bands.

1043 We recommend installing it via the `devtools` package by:

1044 `devtools::install_github("LGraz/CorrectTimeSeries")`

1045 In the following, we shall give a stand-alone example of how the R package can be used:

```
1046 1 library(CorrectTimeSeries)
1047 2
1048 3 # load a list of dataframes, each one describes one pixel with the covariates and
1049 4 # the response
1050 5 data(timeseries_list)
1051 6 str(timeseries_list[[1]])
1052 7
1053 8 # Train/Load RF
1054 9 train_model_myself <- TRUE
1055 10 if (train_model_myself){
1056 11     # Add "true" NDVI (or generally the response), by Out-Of-Bag estimation
1057 12     timeseries_list <- lapply(timeseries_list, function(df) {
1058 13         df$oob_ndvi <- OOB_est(df$gdd, df$ndvi_observed) # gdd is the time-axis
1059 14         df
1060 15     })
1061 16     # Train correction model
1062 17     formula <- "oob_ndvi ~ B02+B03+B04+B05+B06+B07+B08+B8A+B11+B12+scl_class"
1063 18     RF <- train_RF_with_fromula(formula, timeseries_list, robustify=TRUE)
1064 19 } else {
```

```
1066 19  data(RF_for_NDVI)
1067 20  RF <- RF_for_NDVI
1068 21 }
1069 22
1070 23 # ADD CORRECTION
1071 24 timeseries_list <- lapply(timeseries_list, function(df) {
1072 25   df$corrected_ndvi <- randomForest:::predict.randomForest(RF, df)
1073 26   df
1074 27 })
1075 28
1076 29 # Get interpolation for each timeseries
1077 30 newx <- 1:1000
1078 31 lapply(timeseries_list, function(df){
1079 32   ss <- smoothing_spline(df$gdd, df$corrected_ndvi)
1080 33   predict(ss, newx)$y
1081 34 })
```

Example of how to use the `CorrectTimeSeries` package

1083 **Appendix B**

1084 **Further Material**

1085 **B.1 Data and Methods**

1086 **B.1.1 GDD**

1087 Bailey (2018) tabulates the corresponding GDD for each stage of wheat.

Stage	Description	GDD
Emergence	Leaf tip just emerging from above-ground coleoptile.	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

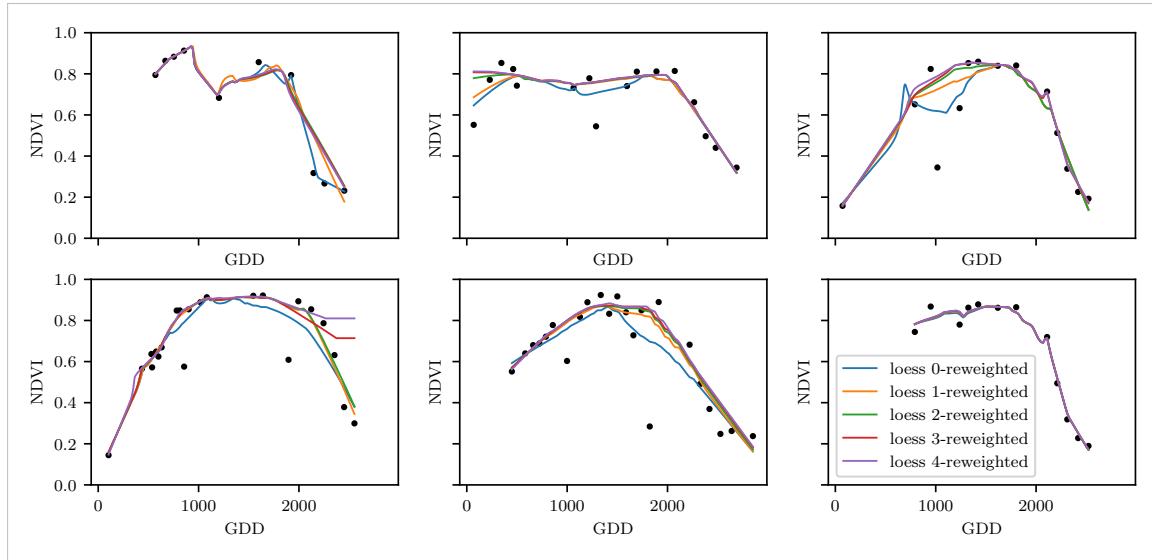
1090 **B.2 Interpolation**

Figure B.1: The LOESS smoother fitted to different (SCL45) NDVI TS. Iterations of a robustifying refit (as indicated in section 3.5) are also displayed.

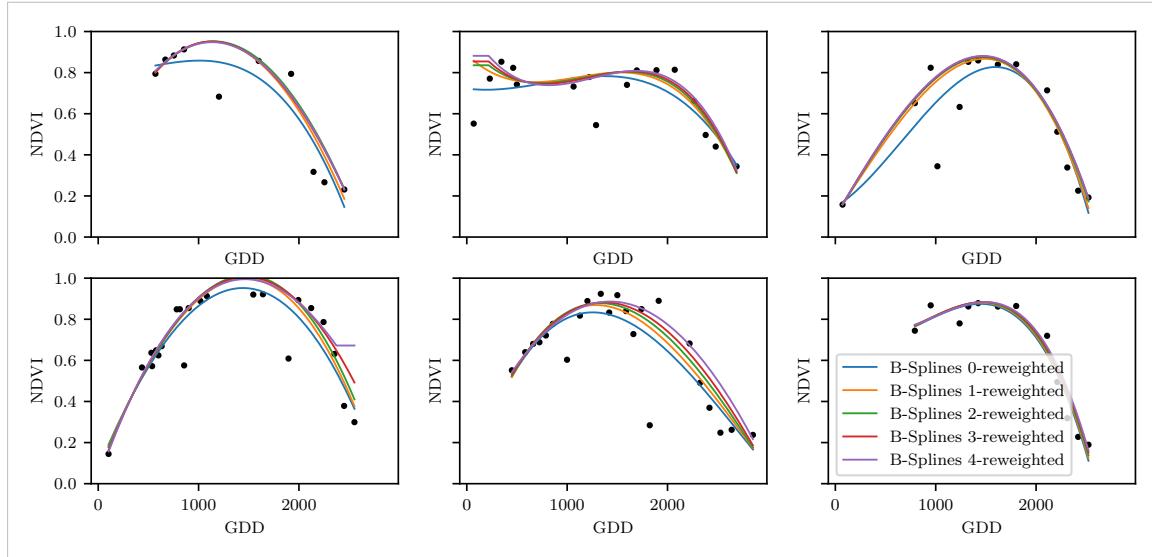


Figure B.2: B-splines fitted to different (SCL45) NDVI TS. Iterations of a robustifying refit (as indicated in section 3.5) are also displayed.

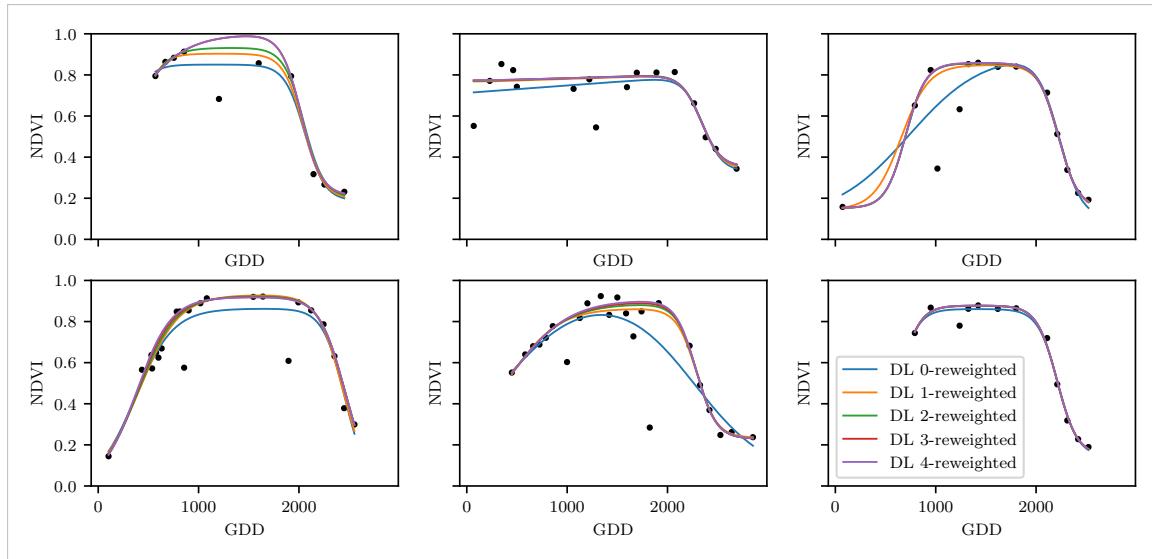


Figure B.3: A Double Logistic curve fitted to different (SCL45) NDVI TS. Iterations of a robustifying refit (as indicated in section 3.5) are also displayed.

### 1091 B.3 NDVI correction

Table B.1: Non-relative RMSE for yield prediction in [t/ha] (cf. table 5.2)

	RF	OLS <sup>SCL</sup>	OLS <sup>all</sup>	MARS	GAM	LASSO	no corrections
SS	1.144	1.033	1.051	1.042	1.046	1.042	1.095
SS <sup>rob</sup>	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 <sup>rob</sup>	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 5.2)

	RF	OLS <sup>SCL</sup>	OLS <sup>all</sup>	MARS	GAM	LASSO	no corrections
SS	0.431	0.486	0.477	0.481	0.479	0.481	0.455
SS <sup>rob</sup>	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 <sup>rob</sup>	0.423	0.439	0.444	0.470	0.456	0.450	0.424

#### 1092 B.3.1 OLS<sup>SCL</sup> Model Outputs

```

1093
1094 1 Call:
1095 lm(formula = (paste(response, " ~ ", "ndvi_observed + scl_class"))),
1096   data = ndvi_df)
1097
1098 2 Residuals:
1099   Min     1Q Median     3Q    Max
1100 -0.7997 -0.0717  0.0039  0.0695  0.6632
1101
1102 3 Coefficients:
1103                               Estimate Std. Error t value Pr(>|t|)
1104 (Intercept)      0.21465     0.00230   93.46 < 2e-16 ***

```

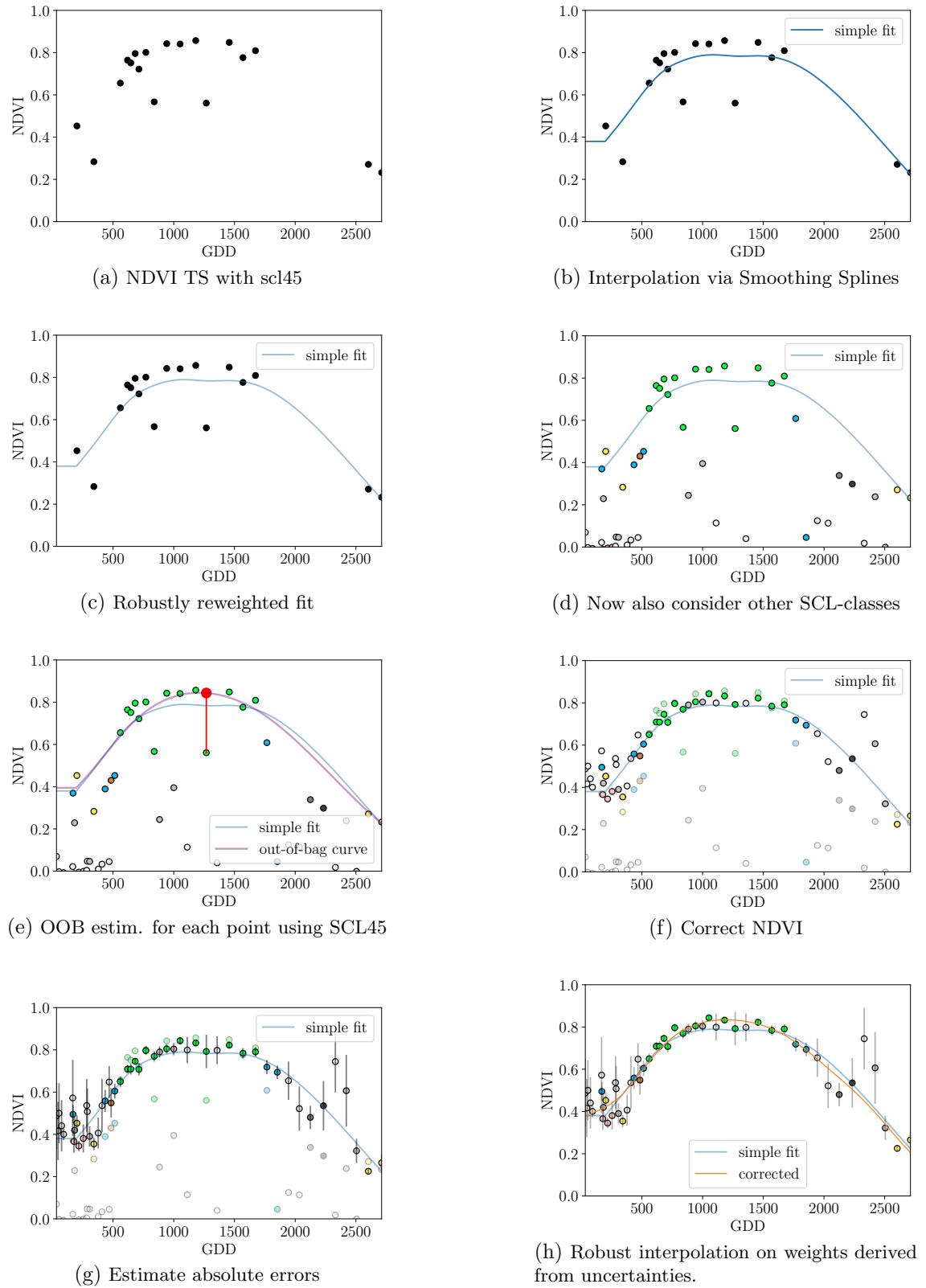


Figure B.4: Stepwise illustration of robust NDVI-Correction. For the color encoding of the SCL classes we refer to table 2.2.

1105	12	ndvi_observed	0.71116	0.00346	205.65	< 2e-16 ***
1106	13	scl_class3	0.02205	0.00356	6.20	5.8e-10 ***
1107	14	scl_class4	-0.00431	0.00251	-1.72	0.085 .

```

1108 15  scl_class5    -0.09875    0.00234   -42.15   < 2e-16 ***  

1109 16  scl_class6    -0.05301    0.01104   -4.80    1.6e-06 ***  

1110 17  scl_class7    0.11245    0.00274   41.09    < 2e-16 ***  

1111 18  scl_class8    0.25963    0.00253   102.57   < 2e-16 ***  

1112 19  scl_class9    0.35994    0.00236   152.47   < 2e-16 ***  

1113 20  scl_class10   0.09091    0.00308   29.54    < 2e-16 ***  

1114 21  scl_class11   0.29784    0.00392   76.06    < 2e-16 ***  

1115 ---  

1116 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  

1117  

1118 Residual standard error: 0.146 on 124978 degrees of freedom  

1119 Multiple R-squared:  0.532,      Adjusted R-squared:  0.532  

1120 F-statistic: 1.42e+04 on 10 and 124978 DF,  p-value: <2e-16

```

R Summary of the NDVI correction model (cf. equation 5.2.0.1)

```

1122  

1123 1 Call:  

1124 2 lm(formula = (paste(get_res(), " ~ ", "ndvi_observed + scl_class")),  

1125 3   data = ndvi_df)  

1126 4  

1127 5 Residuals:  

1128 6   Min     1Q   Median     3Q     Max  

1129 7 -0.2051 -0.0427 -0.0074  0.0329  0.6589  

1130 8  

1131 9 Coefficients:  

1132          Estimate Std. Error t value Pr(>|t|)  

1133 10 (Intercept) 0.18647  0.00126 147.74   < 2e-16 ***  

1134 11 ndvi_observed -0.13265  0.00190 -69.80   < 2e-16 ***  

1135 12 scl_class3 -0.00180  0.00196 -0.92    0.3587  

1136 13 scl_class4 -0.04069  0.00138 -29.55   < 2e-16 ***  

1137 14 scl_class5 -0.09698  0.00129 -75.32   < 2e-16 ***  

1138 15 scl_class6 -0.01906  0.00606 -3.14    0.0017 **  

1139 16 scl_class7  0.01641  0.00150 10.91    < 2e-16 ***  

1140 17 scl_class8 -0.00560  0.00139 -4.02    5.7e-05 ***  

1141 18 scl_class9 -0.01384  0.00130 -10.67   < 2e-16 ***  

1142 19 scl_class10 -0.00690  0.00169 -4.08    4.5e-05 ***  

1143 20 scl_class11 -0.01446  0.00215 -6.72    1.8e-11 ***  

1144 21 ---  

1145 22 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  

1146 23  

1147 24 Residual standard error: 0.08 on 124978 degrees of freedom  

1148 25 Multiple R-squared:  0.352,      Adjusted R-squared:  0.352  

1149 26 F-statistic: 6.8e+03 on 10 and 124978 DF,  p-value: <2e-16

```

R Summary of the NDVI correction model (cf. equation 5.2.0.2)

1151 check quantile und LOOCV definitions

1152 figure spacing (caption zu nah dran — manuell vspace einfügen wo nötig)