



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

**Interpolation and Correction**

9

**of**

10

**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 It is necessary to highlight that without Gregor Perich this project would not have been  
25 possible. His high personal commitment, reliability as well as the weekly instructive su-  
26 pervision meetings were, without 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 more than  
62 5%.

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

65 **Contents**

66	<b>Notation</b>	vi
67	<b>1 Introduction</b>	1
68	<b>2 Data and Methods</b>	3
69	2.1 Sentinel 2 Data . . . . .	3
70	2.2 Crop Yield Data . . . . .	3
71	2.3 Normalized Difference Vegetation Index (NDVI) . . . . .	4
72	2.4 Timescale Transformation . . . . .	5
73	2.5 The Concept of a ‘Pixel’ . . . . .	6
74	2.6 Challenges in S2 Data . . . . .	6
75	2.7 General Methods . . . . .	6
76	2.7.1 Root Mean Square Error (RMSE) . . . . .	8
77	2.7.2 Out-Of-Bag (OOB) and Leave-One-Out-Cross-Validation (LOOCV) . . . . .	8
78	<b>3 Interpolation Methods (IMs)</b>	9
79	3.1 Interpolation Setup . . . . .	9
80	3.2 Parametric Regression . . . . .	9
81	3.2.1 Double Logistic (DL) . . . . .	11
82	3.2.2 Fourier Series (FS) . . . . .	11
83	3.2.3 Optimization Issues . . . . .	12
84	3.3 Non-Parametric Regression . . . . .	12
85	3.3.1 Kernel Regression: Nadaraya-Watson (NW) . . . . .	12
86	3.3.2 Universal Kriging (UK) . . . . .	13
87	3.3.3 Savitzky-Golay Filter (SG) . . . . .	15
88	3.3.4 Locally Weighted Regression (LOESS) . . . . .	16
89	3.3.5 B-Splines (BS) . . . . .	17
90	3.3.6 Smoothing Splines (SS) . . . . .	17
91	3.4 Tuning Parameter Estimation . . . . .	18
92	3.5 Robustification . . . . .	18
93	3.5.1 Our Adjustment: . . . . .	19
94	3.5.2 Examples and Conclusions . . . . .	20
95	3.5.3 Upper Envelope Approach - Penalty for Negative Residuals . . . . .	20
96	3.6 Performance Assessment . . . . .	20
97	<b>4 NDVI Correction XXX.vs.XXX Increase Data Quality</b>	21
98	4.1 Considering other SCL Classes . . . . .	21
99	4.2 Correction Models . . . . .	22
100	4.2.1 Ordinary Least Squares (OLS) . . . . .	23
101	4.2.2 Least Absolute Shrinkage and Selection Operator (LASSO) . . . . .	23
102	4.2.3 General Additive Model (GAM) . . . . .	24
103	4.2.4 Random Forest (RF) . . . . .	24
104	4.2.5 Multivariate Adaptive Regression Splines (MARS) . . . . .	25
105	4.3 Weighted Interpolation . . . . .	26
106	4.4 Resulting Interpolation Strategies (ISs) . . . . .	26
107	4.5 Evaluation via Yield Estimation Accuracy . . . . .	27

108	<b>5 Results</b>	<b>30</b>
109	5.1 Goodness of Fit for Selected IMs . . . . .	30
110	5.2 XXX (Robustification and) NDVI-Correction . . . . .	30
111	<b>6 Discussion</b>	<b>32</b>
112	6.1 IMs . . . . .	32
113	6.1.1 Data Gaps in Time Series . . . . .	32
114	6.1.2 Preselection . . . . .	33
115	6.1.3 Candidate Selection . . . . .	33
116	6.2 NDVI Correction . . . . .	33
117	6.2.1 Choose IS . . . . .	33
118	6.2.2 High RMSE in Yield Prediction . . . . .	33
119	6.2.3 Bootstrap . . . . .	34
120	6.2.4 Using Additional Covariates . . . . .	34
121	<b>7 Conclusion</b>	<b>35</b>
122	7.1 Future Work . . . . .	37
123	7.1.1 Time Series Correction-Interpolation as a General Method . . . . .	37
124	7.1.2 Minor Improvements . . . . .	37
125	<b>Bibliography</b>	<b>38</b>
126	<b>A Reproducibility</b>	<b>41</b>
127	A.1 Reproduce Results . . . . .	41
128	A.2 R-Package . . . . .	41
129	<b>B Further Material</b>	<b>43</b>
130	B.1 Data and Methods . . . . .	43
131	B.1.1 GDD . . . . .	43
132	B.2 Interpolation . . . . .	44
133	B.3 NDVI correction . . . . .	45
134	B.3.1 OLS-SCL Model Outputs . . . . .	45

# 135 Todo list

136	verdeutliche dem leser, dass ein auftrag das findne von interpolationmethoden war . . . . .	9
137	Paper zitieren wo eingeführt oder wo benutzt (falls einföhrung fast schon trivial) . . . . .	9
138	figure / tabelle / pseudocode anstatt aufzählung . . . . .	15
139	consider naming the sub-plots . . . . .	20
140	defition of RYEA, it is not an accuracy but an error . . . . .	30
141	Here in the discussion, you should take up the points you mentioned in the introduction . . . . .	32
142	table mit OLS SCL als sieger diskutieren . . . . .	33
143	kurzer kontext von vergleichbaren values von gregor — diese sektion ist für dena uftraggebenr . . . . .	33
144	even in a perfect world the NDVI curve only holds a fraction of the information avialbe . . . . .	33
145	where does this section belong to? Chapter ‘NDVI Correction’ or ‘Further Work’? . . . . .	34
146	You already capture the ”main” structure of your thesis with the interpolation and the NDVi correction sections. Can you combine them both in a ”synthesis” subsection at the end of the discussion? . . . . .	34
147	Question: more details for the justification of the interpolation candidates? . . . . .	35
148	page breaks . . . . .	45
149	replace space before ref by tilda . . . . .	46
150	check quantile definitions . . . . .	46
151	schwarz weiss färbung der IS tabelle korrigieren . . . . .	46
152	so wenig wie möglich abkürzungen in den fig und table captions . . . . .	46
153	refer to data aviability . . . . .	46
154	abkürzungen Fourier und in tabellen . . . . .	46
155	figure spacing (caption zu nah dran — manuell vspace einfügen wo nötig) . . . . .	46
156	italics für definitionen wie ‘variogramm’ ja/nein — einheitlich . . . . .	47
157	Gross schreiben von Fussnoten & tabelleneinträgen + Satzzeichen . . . . .	47

# 161 Notations

## 162 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 $X$

## 163 Abbreviations and Objects

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.	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 4), (possibly different) IMs (iteratively) used, weightings applied (cf. robustification in section 3.5). Note, that strictly speaking every IM also is an IS. But usually we expect an IS to involve a more ‘complex’ procedure.
IS	

---

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

---

S2	Sentinel 2 satellites. Two multi-spectral image satellites deployed by the European Space Agency.
SCL	Scene Classification Layer provided by the European Space Agency that gives an estimation of the land cover class of each pixel. It indicates what one can expect at a pixel at a sampled time. For an overview, see table 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.
NDVI	Normalized Difference Vegetation Index (Rouse, 1974)
DAS	Days After Sowing
GDD	Growing Degree Days – cumulative sum of “max(0, temperature – threshold)”
RYEA	Relative Yield-Estimation-Accuracy. 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).

## 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-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)

<sup>165</sup> XXX only equations that are referenced are equipped with a number

<sup>166</sup> XXX itpl method and strategy

167 **Chapter 1**

168 **Introduction**

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

191 There are several approaches to adequately address this issue. One is to look at the  
192 observed evolution of vegetation density and assume its bell shape for the NDVI TS given  
193 the strong correlation between NDVI and vegetation density. Approaches to model this  
194 include a 2nd order Fourier approximation ([Stöckli and Vidale, 2004](#)) or a Double Logistic  
195 function ([Beck et al., 2006](#)). On the other hand, assumptions are made about more abstract  
196 properties of the curve, such as smoothness or the like. We divide these into local and  
197 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  
198 to interpolate the TS stepwise. Global methods like B-Splines ([Gurung et al., 2009](#)) and  
199 Smoothing Splines ([Cai et al., 2017](#)) reduce the squares of all residuals simultaneously,  
200 and Universal Kriging fits a Gaussian process to the data ([Chandola and Vatsavai, 2010](#)).

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

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

- 217 i.) Which IMs are used in the context of NDVI and what are their advantages and  
218 disadvantages?
- 219 ii.) How may contaminated data be dealt with?
- 220 iii.) How do data gaps affect interpolation?
- 221 iv.) How to deal with data gaps?
- 222 v.) How can we recognize a good interpolation of the NDVI?

223 The thesis is structured as follows: After presenting the available data, illustrating chal-  
224 lenges and defining different concepts in chapter 2, we turn to the two main blocks of this  
225 thesis. On the first, in section 3 we study parametric and non-parametric IMs (question i.),  
226 generalize an iterative robustification technique (question ii.), and show a way to evaluate  
227 interpolations with out-of-bag residuals (question iii.). In section 6.1.1 we discuss how  
228 different IMs respond to data gaps, and in section 6.1.2 we preselect IMs. We evaluate  
229 this preselection in 5.1 and select two candidates from different IMs in section 6.1.3. For  
230 the second, we attempt to correct contaminated data with statistical models in section  
231 4 (question ii.) and utilize previously ignored observations, which we hope will further  
232 reduce data gaps (question iv.). In addition, we compare different ISs using a vegetation-  
233 oriented quality measure (question v.) and describe the results in section 5.2. Based on  
234 these results, we argue what the best IS is in section 6.2. In addition, we justify why our  
235 NDVI correction can be understood as unsupervised learning and why we relied only on  
236 satellite imagery and not on meterological data for the NDVI correction. Our conclusions  
237 of this thesis, recommendations, as well as an outlook on future work is given in chapter  
238 7.

239 **Chapter 2**

240 **Data and Methods**

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

246 **2.1 Sentinel 2 Data**

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

251 The S2 images contain 12 spectral bands with spatial resolutions up to 10 meters (see 2.1).  
252 Bands with a lower resolution (20 and 60 meters) were upscaled to 10 meter resolution using  
253 cubic interpolation (Perich et al., 2022). In order to decrease the effect of atmospheric  
254 conditions like reflections and scattering, bottom-of-atmosphere, radiometric corrected  
255 Level-2A data was used<sup>1</sup>. The European Space Agency also supplies an algorithm (ESA,  
256 2022a) produces Scene Classification Layer (SCL) where for each location the observed  
257 subject is assigned to one of 11 SCL-classes (cf. table 2.2). In this thesis, we will use  
258 this classification to filter out data points, that we believe to be less informative. That are  
259 all observations which SCL-class does not correspond to vegetation or bare soils (classes  
260 4 and 5). For convenience, we define the set SCL45 as the observations that belong to  
261 SCL-class 4 or 5.

262 **2.2 Crop Yield Data**

263 The crop yield data were collected using a combine harvester. Equipped with GPS, the  
264 harvester drives over the fields and continuously estimates the dry crop yield density in  
265  $t/ha$  (see fig. 2.1a). We take the data set derived in Perich et al. (2022), where error-  
266 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

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

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

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

## 2.3 Normalized Difference Vegetation Index (NDVI)

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

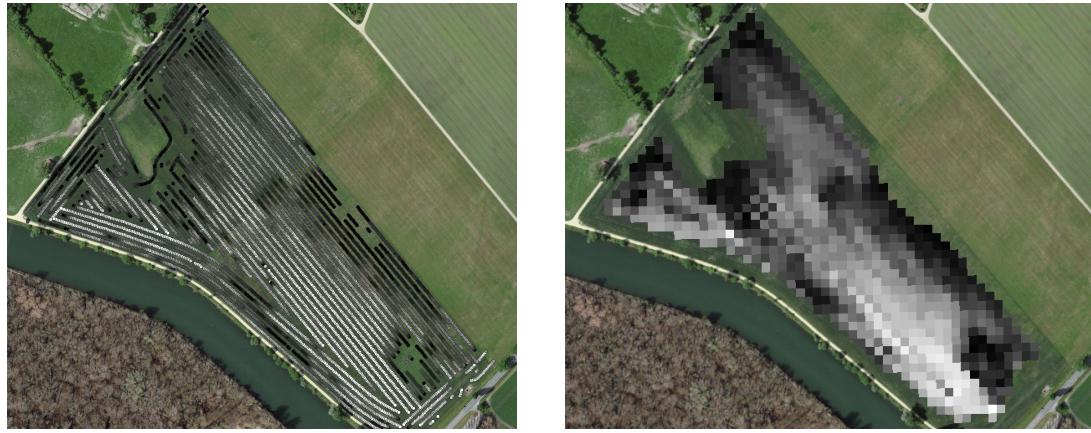


Figure 2.1: Crop yield density map of a field. Ranges from 0.1 t/ha (black) to 5.35 t/ha (white)

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

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

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

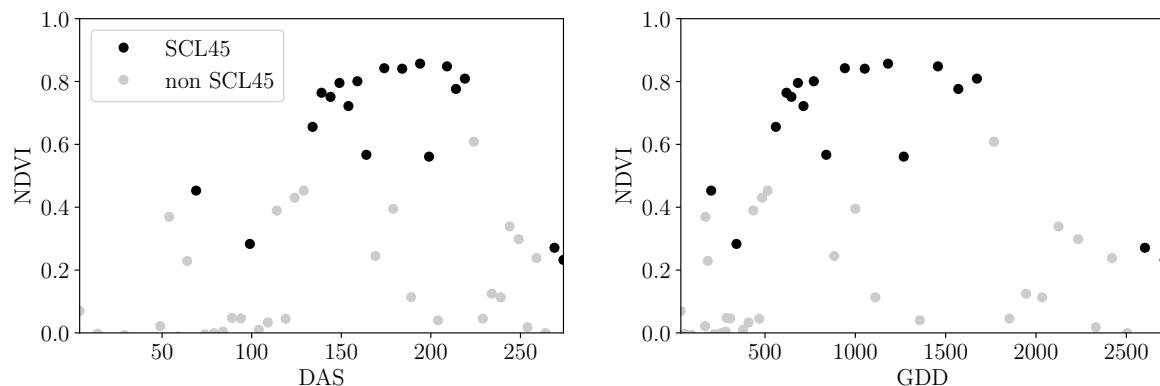


Figure 2.2: NDVI TS plotted against DAS and GDD. GDD are introduced in section 2.4.

284

## 285 2.4 Timescale Transformation

286 Regarding the Days After Sowing (DAS) time scale shown in fig. 2.2a, we detect two  
287 drawbacks. First, this scale makes it difficult to compare two NDVI TS because wheat is  
288 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

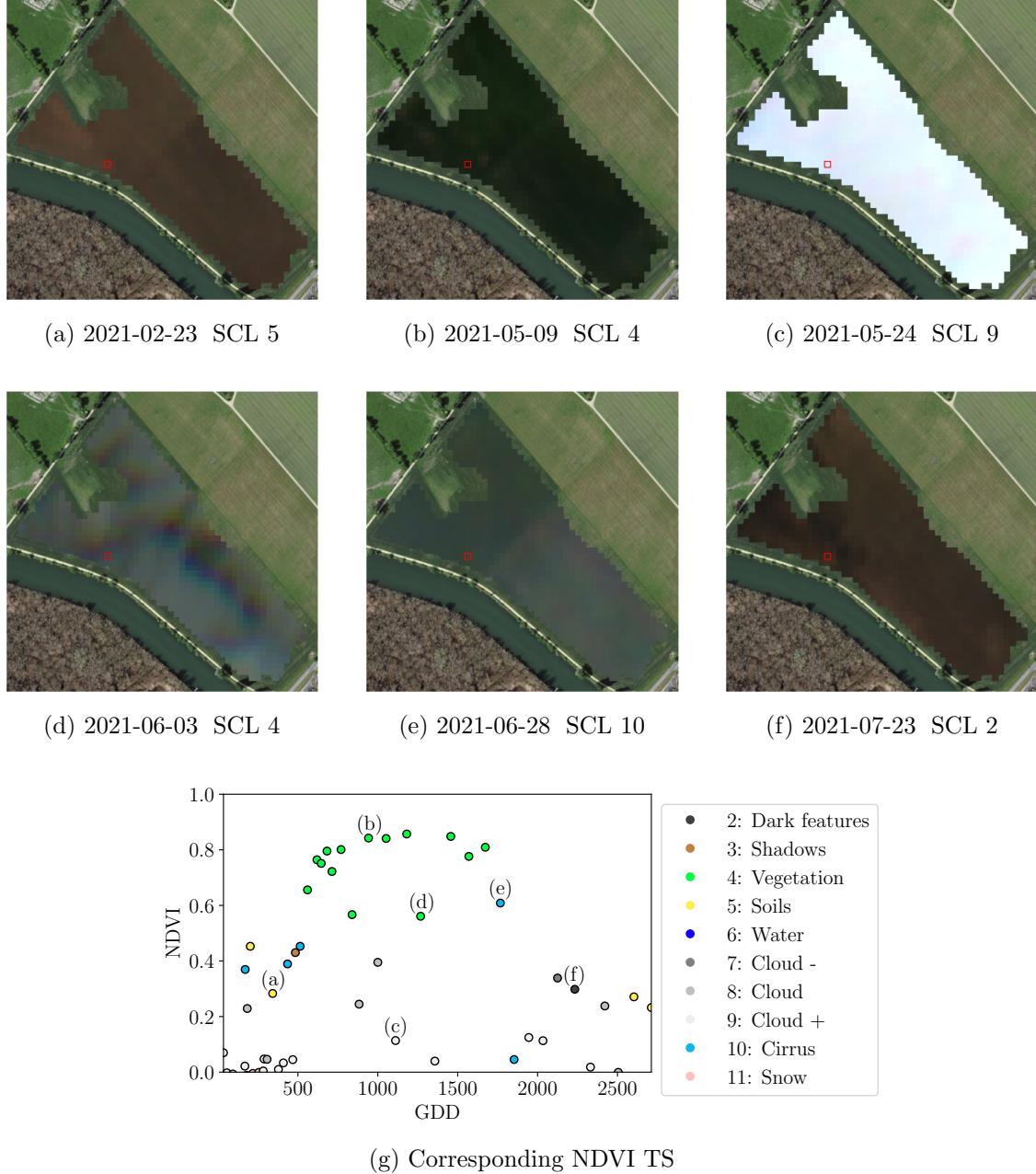


Figure 2.3: Satellite images of a field at selected times with a static background for orientation. Moreover, the NDVI TS of the red-highlighted pixel is shown in (g) colored by the SCL labels.

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

329 **2.7.1 Root Mean Square Error (RMSE)**

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

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

332 **2.7.2 Out-Of-Bag (OOB) and Leave-One-Out-Cross-Validation (LOOCV)**

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

To formally introduce this, let:

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

336 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  
 337 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  
 338 obtain  $\hat{y} := (\hat{y}_1, \dots, \hat{y}_n)$  the OOB estimator for  $y \in \mathbb{R}^n$ .

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

341 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  
 342 the corresponding  $\hat{y}_i$  also the LOOCV estimator.

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

346 **Chapter 3**

347 **Interpolation Methods (IMs)**

348

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

verdeutliche  
dem  
leser,  
dass ein  
auftrag  
das  
findne  
von  
interpo-  
lation-  
metho-  
den war

359 **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

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

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

363 Paper zitieren wo eingeführt oder wo benutzt (falls einföhrung fast schon trivial)

364 **3.2 Parametric Regression**

365 Parametric Curve estimation tries to fit a parametric function, such as, for example, a  
366 Gaussian function with parameters  $\mu$  and  $\sigma$ , to a dataset. In the following, we introduce  
367 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	No

368 **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:

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

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

371 **Robustification**

- 372 Similar as for the SG (cf. section 3.3.3) one can reestimate (only once) the parameters by  
 373 giving less weight to the overestimated observations and more weight to the underestimated  
 374 observations. For the details on the choice of the weights we refer to Beck et al. (2006). We  
 375 will not apply this reestimation but rather the robustification introduced later in section  
 376 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>

377 **3.2.2 Fourier Series (FS)**

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

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

- 378 where  $\Phi = 2\pi \times (t - 1)/n$ . Thus, we periodical behavior. If we would set the period to  
 379 match one year this would coinced with the notion that plans grow every year. Analogous  
 380 to section 3.2.1 we fit it to the data by least squares. Example fits can be seen in figure  
 381 3.1

Advantages	Disadvantages
<ul style="list-style-type: none"> <li>— Assumption of periodicity can be helpful if we are modelling multiyear grow 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>

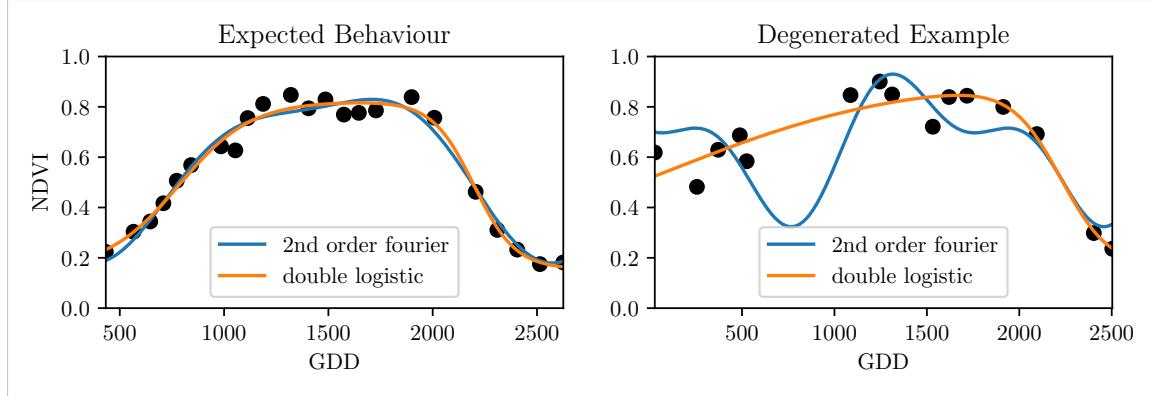


Figure 3.1: Here we observe the possibilities of a precise fit for the two parametric methods but notice also some misbehavior

### 3.2.3 Optimization Issues

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

## 3.3 Non-Parametric Regression

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

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

As described in section 3.1, we aim to estimate

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

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

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

<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:

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

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

### 398 Choose Bandwidth

399 Note that we still need to choose the bandwidth of the function. This can be done with  
 400 the help of LOOCV while optimizing the RMSE. For non-equidistant data we refere to  
 401 Brockmann et al. (1993) where a local adaptive bandwidth selection is presented.

Advantages	Disadvantages
— fletible 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-matrit)	— choice of bandwidth, especially if $t_i$ are not equidistant.
— estimation of the noise variance $\hat{\sigma}_\varepsilon^2$ (REF cf. CompStat 3.2.2)	

### 402 3.3.2 Universal Kriging (UK)

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

407 A Gaussian Process  $\{S(t) : t \in \mathbb{R}\}$  is a stochastic process if  $(S(t_1), \dots, S(t_k))$  has a multi-  
 408 variate Gaussian distribution for every collection of times  $t_1, \dots, t_k$ .  $S$  can be fully charac-  
 409 terized by the mean  $\mu(t) := E[S(t)]$  and its covariance function  $\gamma(t, t') := \text{Cov}(S(t), S(t'))$ .  
 410 Furthermore, we will assume the Gaussian process to be stationary. That is for  $\mu(t)$  to be  
 411 constant in  $t$  and  $\gamma(t, t')$  to depend only on  $h = t - t'$ . Thus, we will write in the following  
 412 only  $\gamma(h)$ .<sup>2</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

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

413 Here  $h$  denotes the distance,  $n$  is the nugget,  $r$  is the range and  $p$  is the partial sill. The  
 414 influence of the parameters is visualized in figure 3.2.<sup>3</sup>

<sup>2</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>3</sup>Strictly speaking we use a scaled version of the variogram. Thus, only the ratio of  $p/n$  matters.

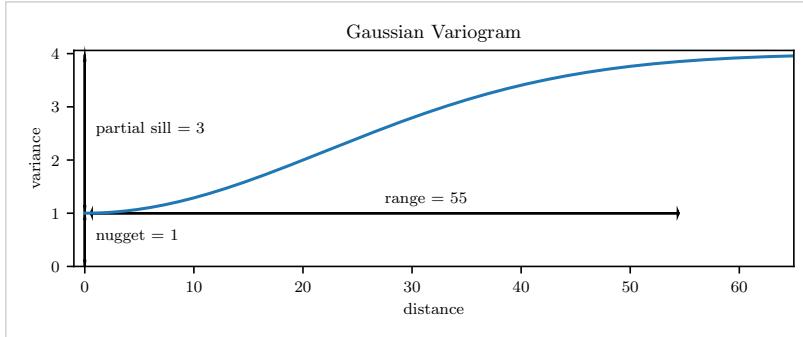


Figure 3.2: Gaussian Variogram with nugget=1, partial sill=3, range=55

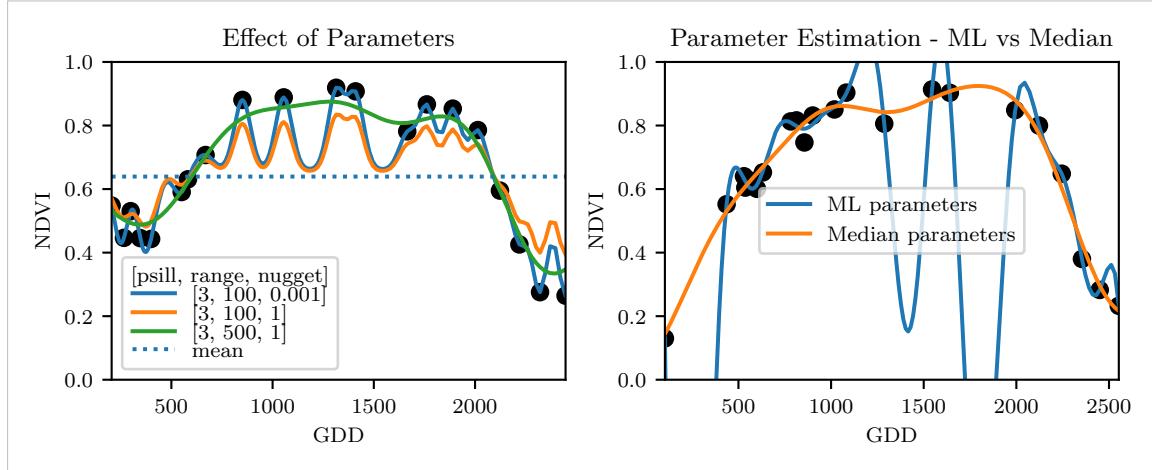


Figure 3.3: On the left, we see how the interpolation change if we increase the nugget and the range parameter. On the right, we compare two UK interpolations, where one takes parameters by numerically maximizing the (which results in a very small nugget) and the other takes the median of many such numerical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

415 Finally, we consider a one-dimensional Gaussian process  $G_\gamma$  with variogram  $\gamma$  and tune the  
416 variogram parameters using maximum likelihood<sup>4</sup>. Let  $z$  be a vector with the new values  
417 to extrapolate, then we can determine the values  $m(z) = \mathbb{E}[G_\gamma(z)|(t, y)]$  using Bayes rule<sup>5</sup>.  
418 For an example fit, we refer to figure 3.3.

#### 419 Violated Assumption

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

<sup>4</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>5</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>

424 **3.3.3 Savitzky-Golay Filter (SG)**

425 The SG, introduced in [Savitzky and Golay \(1964\)](#) is a technique in signal processing and  
 426 can be used to filter out high frequencies (low-pass filter) ([Schafer, 2011](#)). Furthermore,  
 427 it can also be used for smoothing by filtering high frequency noise while keeping the low  
 428 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

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

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

430 [Chen et al. \(2004\)](#) developed a ‘robust’ IM for the NDVI based on the SG. The method  
 431 is based on the assumption that due to atmospheric effects the observed NDVI tends to  
 432 be underestimated and that it cannot increase too quickly. The latter is argued by the  
 433 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>6</sup> and by repeating this step we approach the “upper NDVI envelope”.

figure /  
tabelle /  
pseudocode  
anstatt  
aufzählung

442 **Extension: Spatial-Temporal SG**

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

<sup>6</sup>Here we call a point  $i$  an outlier if  $X_i^0 < X_i^1$ .

445 improve by looking at adjacent pixels<sup>7</sup>. Because we are working with rather high resolution  
 446 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>

447 **3.3.4 Locally Weighted Regression (LOESS)**

448 The LOESS introduced by [Cleveland \(1979\)](#) can be understood as a generalization of the  
 449 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

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

450 where  $h_i$  is the minimal distance such that  $\lceil \alpha n \rceil$  observations are in the ball  $B_{h_i}(t_i)$ .<sup>8</sup> So  
 451 for each  $y_i$  we only consider a proportion  $\alpha$  of the observations.

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

453 The LOESS smoother takes a fraction of points instead of a fixed number and therefore  
 454 automatically adapts to the size of the data we wish to interpolate. However, we run  
 455 into the danger of considering too little observations, since the estimation breaks down if  
 456  $\lceil \alpha n \rceil < d + 1$ .<sup>8</sup> Furthermore, LOESS gives less weight to points further away. This yields a  
 457 "smoother" estimate, since when we slide the window (e.g., for estimating the next value)  
 458 an influential point at the border does not suddenly get zero weight from being weighted  
 459 equally before. Finally, the LOESS also can be used for non-equidistant data and allows  
 460 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>7</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>8</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.

461 **3.3.5 B-Splines (BS)**

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

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

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

463

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

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>9</sup> such that:

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

---

**Advantages**

- can be assigned degrees of freedom
- extendable to "smooth" version
- performs also well if points are not equidistant

---

**Disadvantages**

- smoothing process does not translate well to a interpretation (unlike SS)
  - choice of smoothing parameter  $s$
- 

464 **3.3.6 Smoothing Splines (SS)**

465 Let  $\mathcal{F}$  be the Sobolev space (the space of functions of which the second derivative is  
466 integrable). Then the unique<sup>10</sup> minimizer

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

467 is a cubic spline (i.e., a piecewise cubic polynomial function). The objective function  
468 ensures that we decrease the curvature while keeping the RMSE low.

---

<sup>9</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>10</sup>Strictly speaking it is only unique for  $\lambda > 0$

469 XXX Whittaker

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> </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>

470 

## 3.4 Tuning Parameter Estimation

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

- 475 i.) Construct a set  $\Lambda$  of candidate parameters that generously covers the parameter  
 476 space.
- 477 ii.) Consider  $\mathcal{P}$ , a set of Pixels.
- 478 iii.) For each parameter  $\lambda \in \Lambda$  consider the individual pixels and compute the LOOCV<sup>11</sup>  
 479 for the absolute residuals of the specific NDVI IM for all Pixels in  $\mathcal{P}$  and store them  
 480 in the set  $R_\lambda$ .
- 481 iv.) Determine  $\lambda_{optimal} = \arg \min_{\lambda \in \Lambda} q_{90}(R_\lambda)$ , where we describe the 90% quantile with  
 482  $q_{90}$ .

483 We choose  $\text{quantile}(90)$  as our optimization function because we want to allow 10% of  
 484 outliers (corrupt points) but also aim for an accurate fit in 90% of the cases.

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

487 

## 3.5 Robustification

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

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

<sup>11</sup>For a definition of the leave-one-out-cross-validation we refer to section 2.7.2

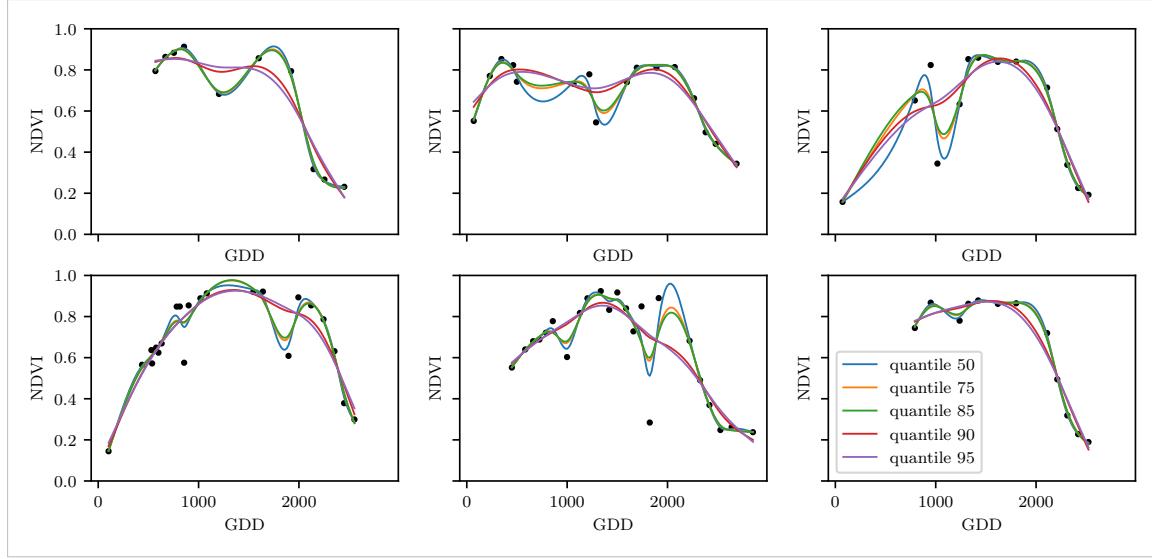


Figure 3.4: Smoothing splines fit with smoothing parameter optimized by minimizing the given quantile of the absolute leave-one-out residuals. Note that the larger the considered quantile is, the smoother the resulting curve becomes.

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

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

496 Next, we compute new weights by

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

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

499 Note that this procedure is indeed robust since we use the median for the normalization  
 500 which has a breakdown point<sup>12</sup> of 50%.<sup>13</sup>

### 501 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

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

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

<sup>12</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>13</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).

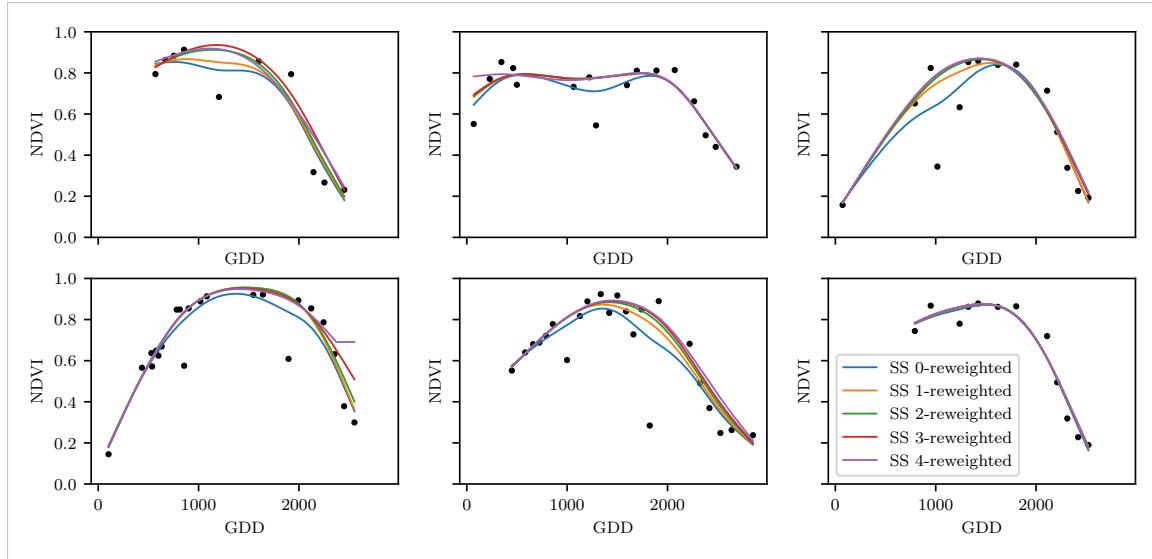
503 **3.5.2 Examples and Conclusions**

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

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

consider naming the subplots

511 **3.5.3 Upper Envelope Approach - Penalty for Negative Residuals**

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

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

519 **3.6 Performance Assessment**

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

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

526 **Chapter 4**

527 **NDVI Correction XXX.vs.XXX**  
528 **Increase Data Quality**

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

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

543 **4.1 Considering other SCL Classes**

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

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

550 **Assumption 1.** The “true” NDVI value at time  $t$  can be successfully estimated by robustified  
551 LOOCV interpolation using high-quality observations. That is, the interpolated value  
552 (using a robustified IM from chapter 3) considering the points  $P^{SCL45} \setminus P_t$ . In the following,  
553 we will call this estimate the “true”-NDVI.

554 We would like to get an idea if there is any information that can be recovered from non-  
555 SCL45 observations. For that, we will check for the other SCL-classes if there is a relation  
556 between the “true” NDVI (derived with robustified SS) and the observed NDVI. Thus, we

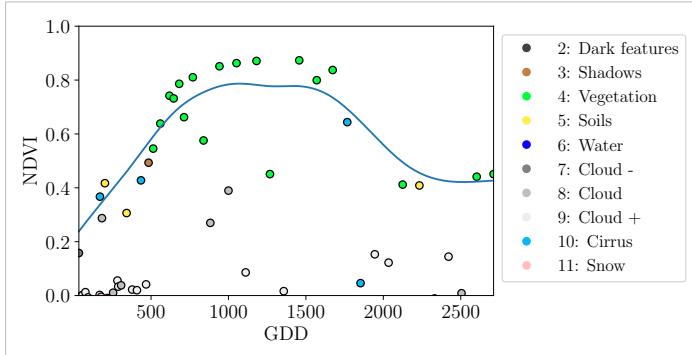


Figure 4.1: A smoothing splines fit considering green and yellow points (SCL45)

pair each “true” NDVI with its observed one, collect all pairs, and create a scatter plot for each SCL-class in fig 4.2. As expected, the “true” and the observed NDVI seem to be highly correlated for SCL45. But we can also detect some patterns of correlation in the SCL-classes 2, 3, 7, 8 and 10.

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

## 4.2 Correction Models

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

In the following, we will introduce different approaches, to model the relationship between the response  $y := \text{NDVI}^{\text{true}} \in \mathbb{R}^n$  and the design matrix  $X \in \mathbb{R}^{n \times p}$ . First, we will study the basic OLS. Second, we look at the LASSO, an penalized adaptation of the OLS which is known to successfully deal with highly correlated covariates. Afterwards, GAMs are introduced which model the response similar to OLS but allow for non-linear relations. Last but not least, we discuss RF and MARS, which are both flexible modelling approaches.

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

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

585 **4.2.1 Ordinary Least Squares (OLS)**

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

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

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

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

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

Advantages	Disadvantages
— Simple method with good interpretability of coefficients.	— Catches only linear relationships.
— Computationally cheap.	— No integrated variable selection. <sup>2</sup>

591 **4.2.2 Least Absolute Shrinkage and Selection Operator (LASSO)**

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

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

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

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

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

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

<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  $\{1, a, b, ab, a^2, b^2\}$ .

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

607 **4.2.3 General Additive Model (GAM)**

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

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

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

- 1)  $\hat{g}_1 = \mathcal{S}_1(y - \mu)$
- 2)  $\hat{g}_j = \mathcal{S}_j(y - \mu - \hat{g}_1(X_{:,1}) - \dots - \hat{g}_{j-1}(X_{:,j-1}))$  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)  $\hat{g}_j = \mathcal{S}_j(y - \mu - \sum_{k \neq j} \hat{g}_k(X_{:,k}))$  for  $j = 2, \dots, p$
- ⋮

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

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

618 **4.2.4 Random Forest (RF)**

619 To define a random Forest introduced by [Breiman \(2001\)](#) we will first define what a Tree  
 620 is. A (*decision*) *Tree* is a graph  $(V, E)$  without circles, a distinct root node, every node  
 621 has at most two children and every leaf has a value assigned to it. At each node there  
 622 is a boolean condition testing if one variable is greater than some value and a pointer to  
 623 one child depending on the boolean value. To evaluate a tree we start at the root node,

<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.

624 test the boolean expression and go to the node indicated by the resulting pointer. This  
 625 we repeat until we end up at a leaf-node, where we return the value assigned to it.

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

629 To build a Random Forest we will bootstrap-aggregate<sup>10</sup> many such Trees<sup>11</sup>. The prediction  
 630 of the Random Forest for a new point  $x$  is then the mean of the predictions from all  
 631 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.

### 632 4.2.5 Multivariate Adaptive Regression Splines (MARS)

633 A MARS model as introduced in Friedman (1991) can be described by

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

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

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

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

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

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

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

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

<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).

<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.

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

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

## 648 4.3 Weighted Interpolation

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

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

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

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

## 662 4.4 Resulting Interpolation Strategies (ISs)

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

- 664 i.) LOOCV Interpolation (+ robustify?) to get “true” NDVI
- 665 ii.) Correction
- 666 iii.) Uncertainty estimation
- 667 iv.) Interpolation (+ robustify?)

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

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

<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.

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

- 675 — We use the same IM each time.  
 676 — Either we robustify both times, or we do not robustify at all.  
 677 — We use the same underlying method for correction and uncertainty estimation.

678 In this fashion, we obtain 28 distinct ISs, which we will benchmark in the next section.

## 679 4.5 Evaluation via Yield Estimation Accuracy

680 In this section, we introduce the relative yield-estimation-accuracy (RYEA) and utilize it  
 681 to evaluate the 28 ISs from section 4.4. The fundamental assumption is that the closer  
 682 the interpolated NDVI TS is to the true one, the better it can be used to determine crop  
 683 yield. Implicitly, we believe that an NDVI TS that better models yield will incorporate  
 684 more true information about the underlying vegetation. Therefore, we want to determine a  
 685 comparable RYEA for each IS and choose it as a benchmark criterion. This is an objective  
 686 measure, since we have not considered crop yield in any of our previous steps. Moreover,  
 687 this criterion is justified by the fact that yield estimation has been a motivation for the  
 688 interpolation.

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

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

692 where  $\bar{y}$  denotes the sample mean.

693 We would like to estimate the yield from the NDVI TS produced by all the ISs for all  
 694 pixels. However, given the high dimensionality and different lengths of the interpolation  
 695 (not every TS has the same start and end point), we must first map each NDVI TS into a  
 696 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

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

701 As a result, for each IS, a matrix is obtained in which each row corresponds to a pixel  
 702 and both the yield and the covariates (computed by applying the above statistics) are

<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. REF

703 contained. Using this matrix, we train a random forest for yield estimation, and compute  
704 the integrated OOB estimates<sup>16</sup>  $\hat{y}$ . Note that the choice of the modeling approach does  
705 not matter much, as long as it is general enough (i.e., able to approximate any function)  
706 and we use the same one for each IS. Finally, for each IS, we calculate the RYEA and  
707 describe 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  $\frac{1}{e}$ ).

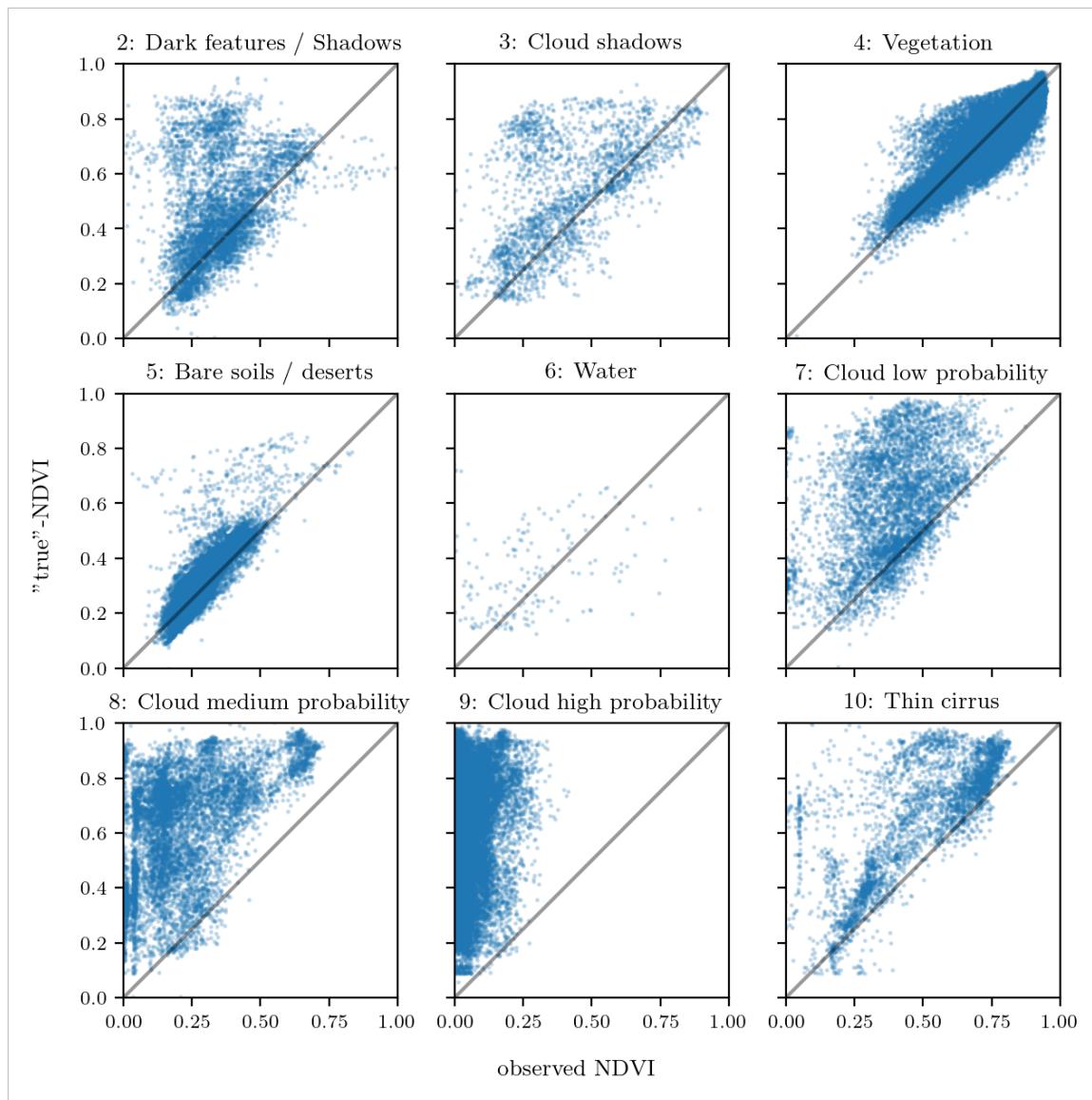


Figure 4.2: For each SCL class, we compare the true NDVI with the observed NDVI. (The true NDVI was estimated with LOOCV smoothing splines, and we used all observations of 10% of the total pixels.)

708 **Chapter 5**

709 **Results**

710 **5.1 Goodness of Fit for Selected IMs**

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

Table 5.1: Comparing the goodness of fit for selectedIMs (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.

	SS	LOESS	DL	BSPL	FR	$SS^{\text{rob}}$	$\text{LOESS}^{\text{rob}}$	$DL^{\text{rob}}$	$BSPL^{\text{rob}}$	$FR^{\text{rob}}$
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

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

721 **5.2 XXX (Robustification and) NDVI-Correction**

722 definition of RYEAs, it is not an accuracy but an error

723 The RYEAs for the 28 (in section 4.4) chosen ISs is given in table 5.2.

724 Robustification in the ISs, does not improve the quality of the fit (measured via the RYEAs)  
725 in most cases. In addition, SS (rob) are better than DL(rob) in terms of RYEAs, with one  
726 exception.

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

Table 5.2: RYEAs. For the non-relative RMSE and the coefficient of determination ( $R^2$ ) see table B.1 and B.2.

	RF	OLS-SCL	OLS-all	MARS	GAM	LASSO	no-correction
ss	0.155	0.140	0.143	0.142	0.142	0.142	0.149
dl	0.156	0.151	0.152	0.152	0.149	0.149	0.158
ss-rob	0.155	0.143	0.147	0.149	0.146	0.145	0.148
dl-rob	0.157	0.153	0.152	0.145	0.148	0.150	0.157

727 The IS that leads to the lowest RYEAs is the OLS-SCL with SS. Given that the OLS-SCL  
 728 models have very good interpretability, we also present the regression equations below.  
 729 The corrected NDVI is calculated using

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

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

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

732 In the equation 5.2.0.1, we notice the strongest upwards correction for SCL classes 8, 9 and  
 733 11 (correspond to ‘medium probability clouds’, ‘high probability clouds’ and ‘thin cirrus  
 734 clouds’). The estimated absolute residuals, however, are the smallest for SCL classes 4 and  
 735 5 (‘vegetation’ and ‘bare soil’). Furthermore, the higher the observed NDVI the lower are  
 736 the estimated absolute residuals.

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

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

739 **Chapter 6**

740 **Discussion**

741 Here in the discussion, you should take up the points you mentioned in the introduction

742 **6.1 IMs**

743 **6.1.1 Data Gaps in Time Series**

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

748 Regarding the GK, we expect that because of the stationarity assumption, the interpolation  
749 will tend to the mean if data gaps are present (cf. figure 3.3).

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

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

759 Similarly, for SS, LOESS, DL and BS we compare the values in table 5.1 between the  
760 robustified and non-robust variant. We find that the robust variant does not differ strongly  
761 from the non-robust variant (unlike FR). Thus, we conclude that these methods do not  
762 have systematic failures.

763 Regarding the LOESS, we observe in the figure B.1 in plot (c) a strange peak between  
764 the first and second observation. This peak is due to the local weighting. In case of data  
765 gaps, the weights can attain non-intuitive values. For instance, the first data point in the  
766 plot, although adjacent to the peak, is given a low weight compared to the points to the  
767 right of the peak (for estimating the value at this peak).

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

### 771 6.1.2 Preselection

772 We shall now justify our preselection of the IMs tested in section 3.6. We decided against  
 773 NW because it has systematic errors at peaks and valleys. Moreover, this method handles  
 774 data gaps poorly (cf. 6.1.1). Moreover, we will not consider UK since the underlying  
 775 assumptions are not met and therefore a systematic bias is introduced. On top of that,  
 776 ML parameter finding occasionally fails. Also, we do not include the SG in the next  
 777 selection, since we think of it as a special case of LOESS.

### 778 6.1.3 Candidate Selection

779 Given that DL convinces regarding most of the selected score functions in table 5.1 we will  
 780 certainly investigate this method in chapter 4. Moreover, we see that the robustification  
 781 mostly improved the score regarding the 50, 75, 85, and 90 % Quantiles. Only for the  
 782 outlier-sensitive score functions (RMSE and q95)<sup>1</sup> we notice significant worsening (we  
 783 consider the robust FS separately in section 6.1.1). Consequently, we will also use the  
 784 robustification in section 4. Not wanting to rely on the form assumptions of the DL, we  
 785 further choose a non-parametric method for further consideration. Despite the LOESS  
 786 slightly dominating the SS in table 5.1, we choose the SS. This is due to the strange  
 787 behavior of the LOESS in case of data gaps (see section 6.1.1) and the good interpretability  
 788 of the SS using the minimization function 3.3.6.1.

789 XXX discuss results from table B.1

## 790 6.2 NDVI Correction

### 791 6.2.1 Choose IS

792 table mit OLS SCL als sieger diskutieren

793 if we use no-correctionXss-rob instead of OLS-SCLXss we loose  $(0.148 - 0.14)/0.148 =$   
 794 5,4% of the information.

### 795 6.2.2 High RMSE in Yield Prediction

797 How much can we expect to get? We have multiple sources of uncertainty in the data:

- 798 i.) Uncertainty in Yield data collected by the combine harvester
- 799 ii.) Uncertainty in Yield data through rasterization
- 800 iii.) Uncertainty in satellite images through “measurement errors” introduced via clouds  
   and other atmospheric effects
- 802 iv.) Uncertainty introduced by interpolating (especially when long data-gaps are present)

803 even in a perfect world the NDVI curve only holds a fraction of the information  
 avialbe

kurzer  
 kontext  
 von  
 vergle-  
 ichbaren  
 values  
 von  
 gregor  
 — diese  
 sektion  
 ist für  
 dena uf-  
 traggeber

<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.

804 **6.2.3 Bootstrap**

805 The question arises if we can build the correction model on the same year as we want to  
806 apply it on. Usually, a similar approach might carry the danger of overfitting. However, we  
807 have not used any ground truth at any point (until the evaluation). Instead, we estimated  
808 the “true” NDVI with the assumption 1 via OOB. Thus, we have bootstrapped our way  
809 out of the problem. Consequently, we reason that we can apply our method to a new  
810 (comparable) dataset and solve the correction again via this bootstrap.

812 **6.2.4 Using Additional Covariates**

813 In section 4.2 we have only used the spectral data (and the observational NDVI calculated  
814 from them) as covariates. Since we have the weather data available (cf. REF-SEC), it  
815 would be a small effort to incorporate it, together with statistics collected from it (i.e.,  
816 GDD or ‘rainfall in the last 30 days’).

817 We decided against using this data, because on the one hand we have the problem that  
818 we have practically too few observations (we observe only 5 years) and we expect the  
819 weather in our study region to be rather homogeneous which is suggested by the fact  
820 that the weather data published by Meteoswiss are for a grid with a resolution of 1 km.  
821 On the other hand, we want the underlying model not to learn improper relationships.  
822 For example, the model might automatically predict a high NDVI for a day in summer  
823 (detected by high GDD / many sunshine hours / high temperature) just because it is  
824 “used” to observing a lot of vegetation in summer. Including temporally (e.g.,  $P_{t-1}$  and  
825  $P_{t+1}$ ) and geographically adjacent pixels would likely improve performance. However, for  
826 simplicity, we omit it here<sup>2</sup>.

where  
does  
this sec-  
tion be-  
long to?  
Chapter  
‘NDVI  
Correc-  
tion’ or  
‘Further  
Work’?

827 You already capture the “main” structure of your thesis with the interpolation and the  
NDVi correction sections. Can you combine them both in a “synthesis” subsection at  
the end of the discussion?

<sup>2</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).

828 **Chapter 7**

829 **Conclusion**

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

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

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

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

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

863 The traditional answer to the question of how to deal with contaminated observations is  
864 that we only consider observations that are labeled as vegetation or bare soil by the SCL  
865 (SCL45). The unreliability of this labeling, however, is illustrated in figure 2.3. Moreover,  
866 filtered observations (non-SCL45) might still contain valuable information (see section  
867 4.1). Therefore, we do not adhere to traditional (SCL) filtration but instead consider all  
868 observations and correct the observed NDVI with uncertainty estimation. For this, we use  
869 statistical models that take additional information such as the remaining spectral bands,  
870 the current SCL label and the observed NDVI into account. But before we interpolate  
871 the corrected NDVI values, we assign a weight to each observation, corresponding to its  
872 uncertainty. The uncertainty is estimated analogously as the NDVI has been corrected. By  
873 combining different IMs (with and without robustification) with various statistical models,  
874 we obtain 28 different Interpolation Strategies (ISs) (see section 4.4). To assess which of  
875 these ISs is best, we assume that the better the IS, the better it allows interpolated  
876 NDVI TS to predict yield. Surprisingly, the best strategy is the one with non-robust  
877 SS and the simplest static model considered, which uses only the observed NDVI and  
878 SCL classification. Let us recapitulate the best IS: First, we estimate the “true” NDVI  
879 (REF) using SS via LOOCV. Then obtain the corrected NDVI using the OLS-SCL model  
880 (cf. equation 5.2.0.1). Subsequently, we estimate the absolute error with the OLS-SCL  
881 model (cf. equation 5.2.0.1) and thereby obtain weights which are supposed to reflect the  
882 reliability of the corrected NDVI (cf. equation 4.3.0.1). Finally, we perform a weighted  
883 interpolation with SS.

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

891 As to the question which IM we recommend, we consider two cases. If one only intends  
892 to fit a curve to the data as precisely as possible, we recommend the robustified DL,  
893 since it minimizes the LOOCV residuals in most cases (cf. table 5.1). In the event  
894 that one requires an interpolation that contains as much information about the plant as  
895 possible, we recommend the SS. This recommendation is especially valid if we traditionally  
896 consider only SCL45 observations without correcting the proposed NDVI. However, we  
897 recommend the abovementioned IS with NDVI correction, because otherwise over 5% of  
898 the information about the vegetation will be lost from the NDVI TS (cf. section 6.2.1).  
899 In light of all the sources of error (cf. section 6.2.2) and the fact that we only consider the  
900 NDVI TS, we consider the 5% to be a solid improvement.<sup>1</sup>

<sup>1</sup>The 5% 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).

901 7.1 Future Work

902 7.1.1 Time Series Correction-Interpolation as a General Method

903 Throughout this thesis, we developed a correction and IM for the NDVI. However, we never  
904 used features of the NDVI. Only the parameter estimated via cross-validation in chapter  
905 3.4 depends on the scale of the TS. For simplicity, we could thus determine the parameter  
906 using Generalized Cross Validation (as [Ripley and Maechler](#) suggest). Therefore, our  
907 approach of interpolation and correction of TS can be applied to arbitrary TS as long as  
908 additional information is available. However, further research is required, to demonstrate  
909 the general usefulness of this approach.

910 Example: Cloud Correction with Uncertainty Estimation and Interpolation

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

916 7.1.2 Minor Improvements

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

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

# 935 Bibliography

- 936 (2007). Gaussian models for geostatistical data. In P. J. Diggle and P. J. Ribeiro (Eds.),  
937 *Model-Based Geostatistics*, pp. 46–78. New York, NY: Springer.
- 938 Bailey, S. J. (2018, July). Using Growing Degree Days to Predict Plant Stages. pp. 8.
- 939 Beck, P. S. A., C. Atzberger, K. A. Høgda, B. Johansen, and A. K. Skidmore (2006,  
940 February). Improved monitoring of vegetation dynamics at very high latitudes: A new  
941 method using MODIS NDVI. *Remote Sensing of Environment* 100(3), 321–334.
- 942 Breiman, L. (2001, October). Random Forests. *Machine Learning* 45(1), 5–32.
- 943 Brockmann, M., T. Gasser, and E. Herrmann (1993, December). Locally Adaptive Band-  
944 width Choice for Kernel Regression Estimators. *Journal of the American Statistical  
945 Association* 88(424), 1302–1309.
- 946 Cai, Z., P. Jönsson, H. Jin, and L. Eklundh (2017, December). Performance of Smoothing  
947 Methods for Reconstructing NDVI Time-Series and Estimating Vegetation Phenology  
948 from MODIS Data. *Remote Sensing* 9(12), 1271.
- 949 Cao, R., Y. Chen, M. Shen, J. Chen, J. Zhou, C. Wang, and W. Yang (2018, Novem-  
950 ber). A simple method to improve the quality of NDVI time-series data by integrating  
951 spatiotemporal information with the Savitzky-Golay filter. *Remote Sensing of Environ-  
952 ment* 217, 244–257.
- 953 Chandola, V. and R. R. Vatsavai (2010). Scalable time series change detection for biomass  
954 monitoring using Gaussian Processes. *Conference on Intelligent Data Understanding*,  
955 14.
- 956 Chen, J., P. Jönsson, M. Tamura, Z. Gu, B. Matsushita, and L. Eklundh (2004, June). A  
957 simple method for reconstructing a high-quality NDVI time-series data set based on the  
958 Savitzky–Golay filter. *Remote Sensing of Environment* 91(3), 332–344.
- 959 Cleveland, W. S. (1979, December). Robust Locally Weighted Regression and Smoothing  
960 Scatterplots. *Journal of the American Statistical Association* 74(368), 829–836.
- 961 Courault, D., L. Hossard, V. Demarez, H. Dechatre, K. Irfan, N. Baghdadi, F. Flamain,  
962 and F. Ruget (2021, July). STICS crop model and Sentinel-2 images for monitoring rice  
963 growth and yield in the Camargue region. *Agronomy for Sustainable Development* 41(4),  
964 49.
- 965 ESA (2022a, August). Level-2A Algorithm Overview.  
966 [https://sentinels.copernicus.eu/web/sentinel/technical-guides/sentinel-2-msi/level-  
967 2a/algorithms](https://sentinels.copernicus.eu/web/sentinel/technical-guides/sentinel-2-msi/level-2a/algorithms).

- 968 ESA (2022b, August). Sentinel-2. <https://sentinel.esa.int/web/sentinel/missions/sentinel-2>.
- 970 Friedman, J. H. (1991, March). Multivariate Adaptive Regression Splines. *The Annals of Statistics* 19(1), 1–67.
- 972 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.
- 975 Hastie, T. and R. Tibshirani (1987, June). Generalized Additive Models: Some Applications. *Journal of the American Statistical Association* 82(398), 371–386.
- 977 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.
- 980 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.
- 983 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.
- 986 Lyche, T. and K. Mørken (2005, January). Spline Methods.
- 987 McMaster, G. S. and W. W. Wilhelm (1997, December). Growing degree-days: One equation, two interpretations. *Agricultural and Forest Meteorology* 87(4), 291–300.
- 989 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.
- 992 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.
- 995 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.
- 998 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.
- 1001 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>.
- 1003 Rouse, J. W. (1974, May). Monitoring the vernal advancement and retrogradation (green wave effect) of natural vegetation. Technical Report NASA-CR-139243.
- 1005 Savitzky, A. and M. J. E. Golay (1964, July). Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry* 36(8), 1627–1639.
- 1007 Schafer, R. W. (2011, July). What Is a Savitzky-Golay Filter? [Lecture Notes]. *IEEE Signal Processing Magazine* 28(4), 111–117.

- 1009 Stephen, M. (2021, July). Earth: Multivariate Adaptive Regression Splines.
- 1010 Stöckli, R. and P. L. Vidale (2004, September). European plant phenology and climate  
1011 as seen in a 20-year AVHRR land-surface parameter dataset. *International Journal of  
1012 Remote Sensing* 25(17), 3303–3330.
- 1013 Strbac, O., M. Milanovic, and V. Ogrizovic (2017, July). Estimation the evapotrasnpiration  
1014 of urban parks with field based and remotely sensed datasets. pp. 13.
- 1015 Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: A retrospective.  
1016 *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73(3), 273–  
1017 282.

1018 **Appendix A**

1019 **Reproducibility**

1020 **A.1 Reproduce Results**

1021 For reproducibility of the whole computations, we refer to our codebase at:

1022 <https://github.com/LGraz/MasterThesis-Code>

1023 In order to reproduce our computations and results, set up the directory as described  
1024 in the README and execute the computations via `./shell_scripts/reproduce.sh`  
1025 and do not execute the python and R scripts by hand (unless you follow the order in  
1026 `./shell_scripts/reproduce.sh`).

1027 **A.2 R-Package**

1028 We also provide an R package for a general time series correction and interpolation if  
1029 additional data is available at:

1030 <https://github.com/LGraz/CorrectTimeSeries>

1031 In our case we consider the NDVI time series and the additional data consists of the unused  
1032 spectral bands.

1033 We recommend installing it via the `devtools` package by:

1034 `devtools::install_github("LGraz/CorrectTimeSeries")`

1035 In the following, we shall give a stand-alone example of how the R package can be used:

```
1036 1 library(CorrectTimeSeries)
1037 2
1038 3 # load a list of dataframes, each one describes one pixel with the covariates and
1039 4 # the response
1040 5 data(timeseries_list)
1041 6 str(timeseries_list[[1]])
1042 7
1043 8 # Train/Load RF
1044 9 train_model_myself <- TRUE
1045 10 if (train_model_myself){
1046 11   # Add "true" NDVI (or generally the response), by Out-Of-Bag estimation
1047 12   timeseries_list <- lapply(timeseries_list, function(df) {
1048 13     df$oob_ndvi <- OOB_est(df$gdd, df$ndvi_observed) # gdd is the time-axis
1049 14     df
1050 15   })
1051 16   # Train correction model
1052 17   formula <- "oob_ndvi ~ B02+B03+B04+B05+B06+B07+B08+B8A+B11+B12+scl_class"
1053 18   RF <- train_RF_with_fromula(formula, timeseries_list, robustify=TRUE)
1054 19 } else {
```

```
1056 19  data(RF_for_NDVI)
1057 20  RF <- RF_for_NDVI
1058 21 }
1059 22
1060 23 # ADD CORRECTION
1061 24 timeseries_list <- lapply(timeseries_list, function(df) {
1062 25   df$corrected_ndvi <- randomForest:::predict.randomForest(RF, df)
1063 26   df
1064 27 })
1065 28
1066 29 # Get interpolation for each timeseries
1067 30 newx <- 1:1000
1068 31 lapply(timeseries_list, function(df){
1069 32   ss <- smoothing_spline(df$gdd, df$corrected_ndvi)
1070 33   predict(ss, newx)$y
1071 34 })
```

Example of how to use the `CorrectTimeSeries` package

1073 **Appendix B**

1074 **Further Material**

1075 **B.1 Data and Methods**

1076 **B.1.1 GDD**

1077 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

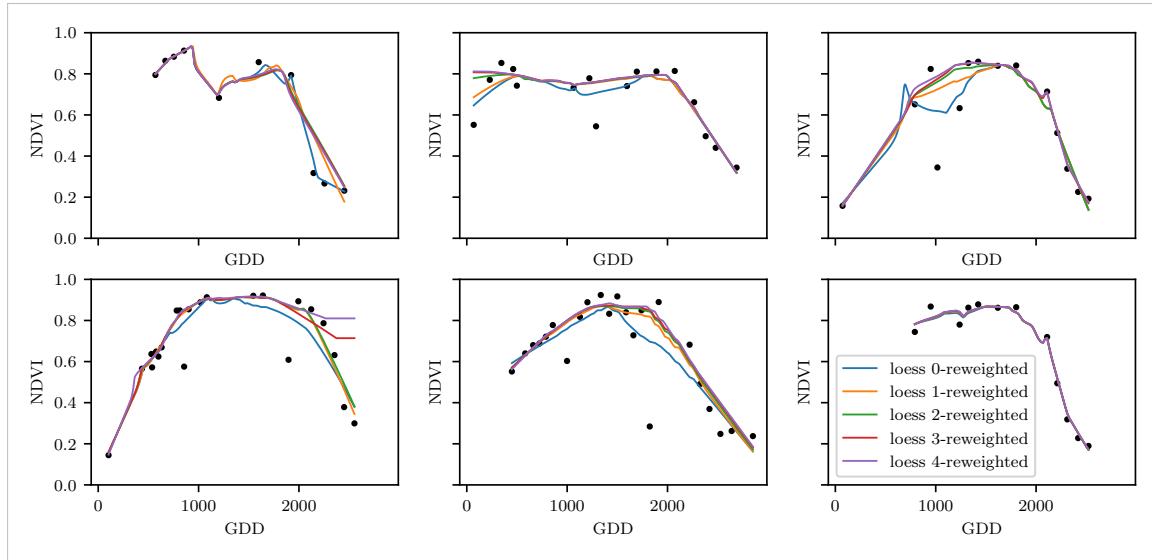
1078 **B.2 Interpolation**

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

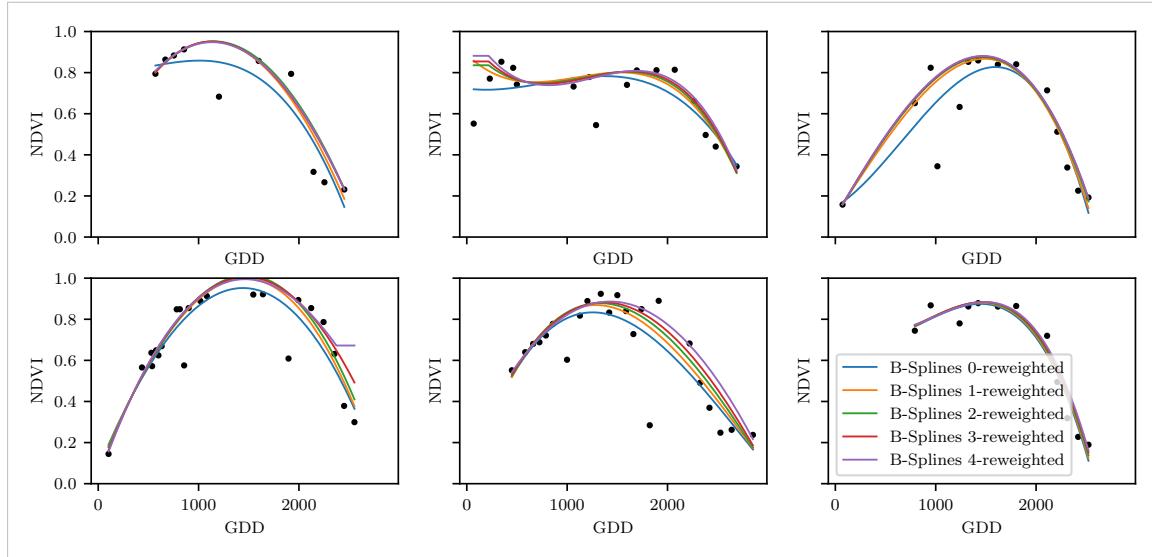


Figure B.2: B-splines fitted to different (SCL45) NDVI TS. Iterations of a robustifying refit (as indicated in section 3.5) are also displayed

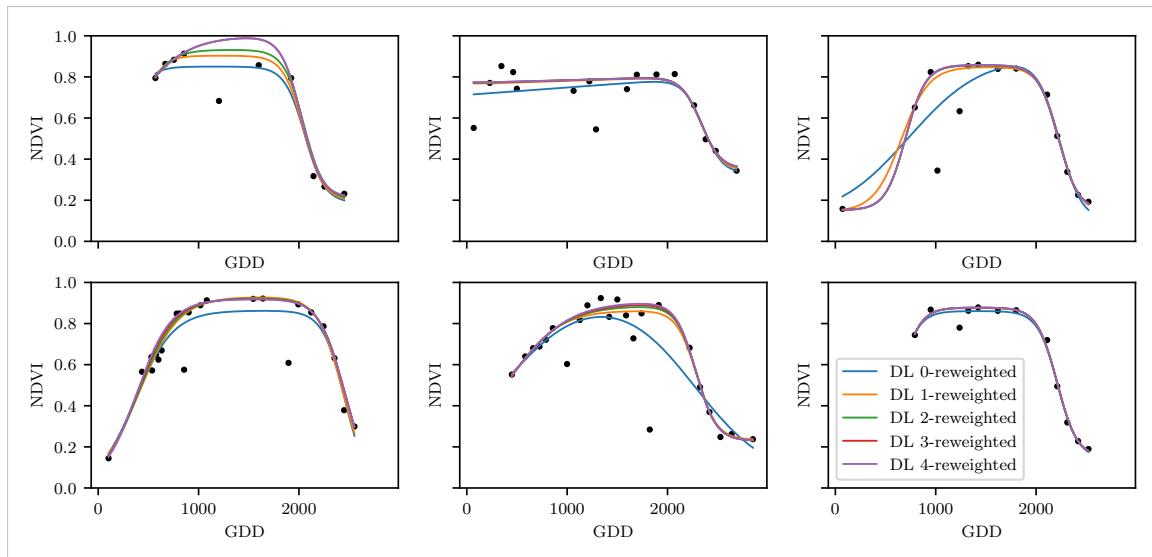


Figure B.3: A Double Logistic curve fitted to different (SCL45) NDVI TS. Iterations of a robustifying refit (as indicated in section 3.5) are also displayed

### 1079 B.3 NDVI correction

1080 page breaks

Table B.1: Non-relative RMSE for yield prediction

	RF	OLS-SCL	OLS-all	MARS	GAM	LASSO	no-correction
ss	1.144	1.033	1.051	1.042	1.046	1.042	1.095
dl	1.150	1.115	1.116	1.116	1.097	1.098	1.159
ss-rob	1.144	1.054	1.084	1.094	1.072	1.071	1.091
dl-rob	1.159	1.128	1.117	1.064	1.093	1.105	1.156

Table B.2: Coefficient of determination ( $R^2$ ) of yield prediction

	RF	OLS-SCL	OLS-all	MARS	GAM	LASSO	no-correction
ss	0.431	0.486	0.477	0.481	0.479	0.481	0.455
dl	0.427	0.445	0.444	0.444	0.454	0.453	0.423
ss-rob	0.431	0.475	0.461	0.456	0.467	0.467	0.457
dl-rob	0.423	0.439	0.444	0.470	0.456	0.450	0.424

#### 1081 B.3.1 OLS-SCL Model Outputs

```

1 Call:
2 lm(formula = (paste(response, " ~ ", "ndvi_observed + scl_class"))),
3   data = ndvi_df)
4
5 Residuals:
6   Min     1Q Median     3Q    Max
7 -0.7997 -0.0717  0.0039  0.0695  0.6632
8
9 Coefficients:

```

```

1092      Estimate Std. Error t value Pr(>|t|)
1093 (Intercept) 0.21465  0.00230  93.46 < 2e-16 ***
1094 ndvi_observed 0.71116  0.00346 205.65 < 2e-16 ***
1095 scl_class3 0.02205  0.00356   6.20  5.8e-10 ***
1096 scl_class4 -0.00431  0.00251  -1.72   0.085 .
1097 scl_class5 -0.09875  0.00234 -42.15 < 2e-16 ***
1098 scl_class6 -0.05301  0.01104  -4.80  1.6e-06 ***
1099 scl_class7 0.11245  0.00274  41.09 < 2e-16 ***
1100 scl_class8 0.25963  0.00253 102.57 < 2e-16 ***
1101 scl_class9 0.35994  0.00236 152.47 < 2e-16 ***
1102 scl_class10 0.09091  0.00308  29.54 < 2e-16 ***
1103 scl_class11 0.29784  0.00392  76.06 < 2e-16 ***
1104 ---
1105 Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ',' 1
1106
1107 Residual standard error: 0.146 on 124978 degrees of freedom
1108 Multiple R-squared: 0.532, Adjusted R-squared: 0.532
1109 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)

```

1111
1112 Call:
1113 lm(formula = (paste(get_res(), " ~ ", "ndvi_observed + scl_class"))),
1114   data = ndvi_df)
1115
1116 Residuals:
1117   Min     1Q   Median     3Q    Max
1118 -0.2051 -0.0427 -0.0074  0.0329  0.6589
1119
1120 Coefficients:
1121             Estimate Std. Error t value Pr(>|t|)
1122 (Intercept) 0.18647  0.00126 147.74 < 2e-16 ***
1123 ndvi_observed -0.13265  0.00190 -69.80 < 2e-16 ***
1124 scl_class3 -0.00180  0.00196  -0.92  0.3587
1125 scl_class4 -0.04069  0.00138 -29.55 < 2e-16 ***
1126 scl_class5 -0.09698  0.00129 -75.32 < 2e-16 ***
1127 scl_class6 -0.01906  0.00606  -3.14  0.0017 **
1128 scl_class7 0.01641  0.00150  10.91 < 2e-16 ***
1129 scl_class8 -0.00560  0.00139  -4.02 5.7e-05 ***
1130 scl_class9 -0.01384  0.00130 -10.67 < 2e-16 ***
1131 scl_class10 -0.00690  0.00169  -4.08 4.5e-05 ***
1132 scl_class11 -0.01446  0.00215  -6.72 1.8e-11 ***
1133 ---
1134 Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ',' 1
1135
1136 Residual standard error: 0.08 on 124978 degrees of freedom
1137 Multiple R-squared: 0.352, Adjusted R-squared: 0.352
1138 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)

1140 replace space before ref by tilda

1141 check quantile definitions

1142 schwarz weiss färbung der IS tabelle korrigieren

1143 so wenig wie möglich abkürzungen in den fig und table captions

1144 refer to data availability

1145 abkürzungen Fourier und in tabellen

1146

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

1147

italics für definitionen wie ‘variogramm’ ja/nein — einheitlich

1148

Gross schreiben von Fussnoten & tabelleneinträgen + Satzzeichen

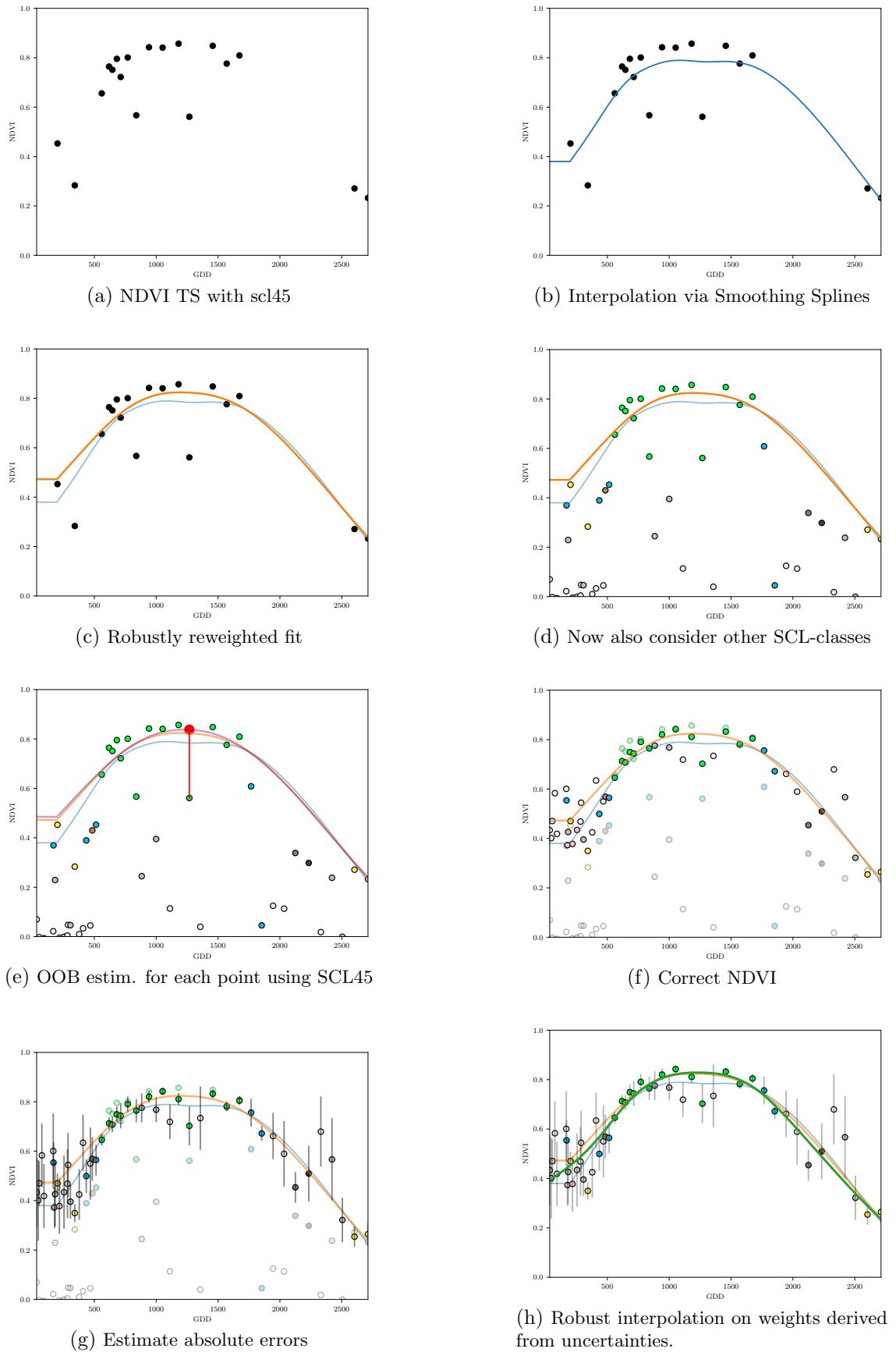


Figure B.4: Stepwise illustration of robust NDVI-Correction. For the color encoding of the SCL classes we refer to table 2.2.