# Contents

# Chapter 1

# Interpolation Methods

In this section we take a closer look at several interpolation methods, which will be used to interpolate and smooth the NDVI timeseries.

First, we give an brief overview in table 1.1.

Second, we introduce and discuss each method.

Then, we try to extract the main ingrediendts of each method to forge our own one.

Finally, using leave-one-out cross validation we tune the parameters (where necessary) and get a first idea of the perfomance of each method.

## 1.1 Methods - Description

### Setting

We are given data in the form of $(x_i, Y_i)$ for $i = 1, \ldots, n)$. Assume that it can be represented by

$$Y_i = m\left(x_i\right) + \varepsilon_i,$$

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

$$m(x) = \mathbb{E}[Y \mid x]$$

Different assumptions on $m$ will lead to the following methods:

### 1.1.1 Kernel Regression

As described previously, we would like to estimate

$$\mathbb{E}[Y \mid X = x] = \int_{\mathbb{R}} y f_{Y|X}(y \mid x) dy = \frac{\int_{\mathbb{R}} y f_{X,Y}(x, y) dy}{f_X(x)}, \tag{1.1.1.1}$$

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

$$\hat{f}_X(x) = \frac{\sum_{i=1}^n K\left(\frac{x - x_i}{h}\right)}{nh}, \hat{f}_{X,Y}(x, y) = \frac{\sum_{i=1}^n K\left(\frac{x - x_i}{h}\right) K\left(\frac{y - Y_i}{h}\right)}{nh^2}$$

| | assumtpions | pros | cons | weights | bounded |
|---|---|---|---|---|---|
| Savitzky-Golay filter | – high frequencies are noise (low.pass filter)<br>– equidistant points<br>– local polynomials | – computationally very fast | – cannot deal natively with missing data (need some interpolation) | no | mostly |
| SG + NDVI | – upper envelope<br>– vegetation cannot grow faster than some slope | – biological knowledge | – bad "upper envelope" since weights are not used for the estimation itselfe | (no) | mostly |
| Loess | – local polynomial with points closer to the estimated point are more important | – flexible<br>– generalization of SG<br>– weighting function makes intuitive sense | – computationally expensive | yes | mostly |
| Smoothing–Splines | 2cd derivative of function is integrable | – intuitive meaning of penalty<br>– general assumptions<br>– flexible shape | – unbounded | yes | no |
| B-Splines (Smoothed) | – function can be approximated by a linear combination of B-splines basis functions | – general assumtion<br>– flexible shape | – unbounded<br>– no intuitive meaning for smoothing | | no |
| (Gaussian) Kernel Smoothing | | – simple<br>– general assumptions | – bandwidh: failes if there are big data-gaps | yes | yes |
| Double-Logistic | – function first increases then decreases<br>– ndvi has a minimal value | – good for evergreen plants (if snow masks ndvi)<br>– upper envelope | – parameterestimation can go seriously wrong<br>– strange behaviour for long data-gaps | yes | mostly |
| Universal Kriging | – function is a realization of a stationary gaussian process | – informative parameters<br>– flexible | – regression to the mean<br>– assumptions clearly not met | yes | mostly |

Table 1.1: A short summary of the studied interpolation methods

By plugging the above into equation 1.1.1.1 we arrive at the *Nadaraya-Watson* kernel estimator:

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

| Pros | Cons |
|---|---|
| — flexible due to different possible kernels<br>— can be assigned degrees of freedom (trace of the hat-matrix)<br>— estimation of the noise variance $\hat{\sigma}_\varepsilon^2$ (XXX c.f. CompStat 3.2.2) | — if the $x \mapsto K(x)$ is not continuous, $\hat{m}$ isn't either<br>— choice of bandwidth, especially if $x_i$ are not equidistant. |

**Examples:** Normal, Box For local bandwidth selection see Brockmann et al. (1993) XXX

### 1.1.2 Savitzky-Golay Filter (SG Filter)

The *Savitzky-Golay Filter*, introduced in Savitzky and Golay (Savitzky and Golay) is a technique in signal processing and can be used to filter out high frequencies (low-pass filter) as argued in Schafer (Schafer). Furthermore, it also can be used for smoothing by filtering high frequency noise while keeping the low frequency signal. First we choose a window size $m$. Then, for each point $j \in \{m, m+1, \ldots, n-m\}$ we fit a polynomial of degree $k$ by:

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

were $P_k$ denotes the Polynomials of degree $k$ over $\mathbb{R}$.

For equidistant points this can efficiently be calculated by

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

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

#### Adaptation to the NDVI

In a rather famous paper Chen, Jönsson, Tamura, Gu, Matsushita, and Eklundh (Chen et al.) a "robust" method based on the Savitzky-Golay has been used. The method is based on the assumption that due to atmospheric effects the observed NDVI tends to be underestimated and that it cannot increase too quickly[1].

**Algorithm:**

i.) Remove points which are labeled as cloudy

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

iii.) Linearly interpolate to obtain an equidistant time series $X^0$

iv.) Apply the Savitzky-Golay Filter to obtain a new time series $X^1$

[1] The latter is argued by the biological impossibility of such fast vegetation changes

v.) Update $X^1$ by applying again a Savitzky-Golay Filter. 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[2] and by repeating this step we approach the "upper NDVI envelope".

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

**Extension: Spatial-Temporal-Savitzky-Golay Filter**

One notable adaptation of the Savitzky-Golay is the presented by Cao, Chen, Shen, Chen, Zhou, Wang, and Yang (Cao et al.). The key difference is the additional assumption of the cloud cover being discontinuous and that we can improve by looking at adjacent pixels[3]. Because we are working with rather high resolution satellite data, and we need the variance in the predictors we will waive this extension.

### 1.1.3 Locally Weighted Regression (LOESS)

Introduced by : Cleveland (Cleveland) implemented here Cappellari, McDermid, Alatalo, Blitz, Bois, Bournaud, Bureau, Crocker, Davies, Davis, de Zeeuw, Duc, Emsellem, Khochfar, Krajnović, Kuntschner, Morganti, Naab, Oosterloo, Sarzi, Scott, Serra, Weijmans, and Young (Cappellari et al.)

The Locally Weighted Regression (LOESS) can be understood as a generlalization of the Savitzky-Golay Filter (c.f. sec. 1.1.2).

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

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

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

**How does the Robust LOESS differ from the SG Filter?**

The Loess smoother takes a fraction of points instead of a fixed number and therefore automatically adapts to the size of the data we wish to interpolate. However, we run

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

[3] 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

into the danger of considering to little observations since the estimation breaks down if $\lceil \alpha n \rceil < d + 1$. Furthermore, Loess gives less weight to points further away. This yields a "smoother" estimate, since when we slide the window (e.g. for estimating the next value) an influential point at the border does not suddenly get zero weight from being weighted equally before. Finally, the Loess also can be used for non-equidistant data and allows for arbitrary interpolation.

**Robustify**

Cleveland (Cleveland) also propose a method how to reduce the impact of outliers.[4]

First we shall introduce a robust measure of the scale (i.e. a robust version of the standard deviation). Define the median of absulute deviation (mad) for $r \in \mathbb{R}^n$ as:

$$\operatorname{mad}(r) := \operatorname{med}\left(|r_i - \operatorname{med}(r)|_{i \in \{1,\dots,n\}}\right) \tag{1.1.3.1}$$

and the bisquare function B:

$$B(x) := \begin{cases} \left(1 - x^2\right)^2, & \text{if } |x| < 1 \\ 0, & \text{else} \end{cases}$$

Now, we do something similar to what is done in iteratively reweighted least squares. Simply, reiterate with updating the weights of each observation with

$$w_i^{\text{new}} := w_i^{\text{old}}\, \mathrm{B}\left(\frac{|r_i|}{6 \operatorname{mad}(r_1, \dots, r_n)}\right) \tag{1.1.3.2}$$

where $r_i = y_i - \hat{y}_i$ denotes the resiudals. Stop the reiteration after several steps or till the change of the values is smaller than some tolerance.

**Our Adjustment:** Since we usually observe outliers with negative residuals we decide to devide the negative residuals by two before updating the weights. Furthermore, we want to prevent low-weighted observations to corrupt our estimation of scale (the mad) and thus we use the weighted mad, where we just use a weighted version of the median. This can be defined as

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

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

[5]

| Pros | Cons |
|---|---|
| — Flexible generalization of Savitzky-Golay | — The nature of local regression might lead to suprising estimates (no smoothness guarantees for the second derivative) |
| — arbitrary interpolation possible | — Multiple XXXXXXx |
| — Intuitive parameters | |

[4] Note that due to using the median for the normalization we gain a breakdown point of 50% for outliers in $y$.

[5] A rather technical adjustment is that we give a reasonable minimal threshold to our scale estimation s.t. the weight $w_i$ wont be set to zero if $r_i < 0.1$. This is done to avoid ignoring to many observations, which could lead to convergence issues or singular matrixes

### 1.1.4 Double Logistic

The Double Logistic smoothing as described in Beck, Atzberger, Høgda, Johansen, and Skidmore (Beck et al.) heavily relies on shape assumptions of the fitted curve (i.e. the NDVI time series).

Assumptions:

— There is a minimum NDVI level $Y_{\text{min}}$ in the winter (e.g. due to evergreen plants), which might be masked by snow. This can be estimated beforehand, taking into several years into account.

— The growth cycle can be divided into an increase and a decrease period where the time series 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:

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

Where the five free parameters: $Y_{\text{max}}$, $d_0$, $d_1$, $t_0$, $t_1$ are initially estimated by least squares. Such fit can be seen in figure 1.1.

Similar as for the Savitzky-Golay Filter (c.f. section 1.1.2) we reestimate (only once) the parameters by giving less weight to the overestimated observations and more weight to the underestimated observations[6].

| Pros | Cons |
|---|---|
| — Incorperates subject specific knowledge in the case of evergreen plants covered by snow. | — Strong shape assumptions on the NDVI curve. |
| — Optimized parameters have an intuitive meaning. | — Parameter optimization might go wrong. This can be mitigated to some extend to provide bounds for the paramters |
| | — Strange behaviour in regrions with little observations. (cf. figure 1.1) |

### 1.1.5 Fourier Approximation

Similar as in section 1.1.4 we fit a parametric curve to the data by least squares. Here we take the second order fourier series:

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

where $\Phi = 2\pi \times (t-1)/n$.

[6] For the details on the weights we refer to Beck, Atzberger, Høgda, Johansen, and Skidmore (Beck et al.)

| Pros | Cons |
|---|---|
| — Assumtion of periodicity can be helpful if we are modelling multiyear grow cycles<br>— Flexible curve shape | — Bad behaviour in regions with little data (cf. figure 1.1)<br>— Hard to interpret estimated parameters<br>— Parameter estimation can go wrong. Introducing bounds can help. |

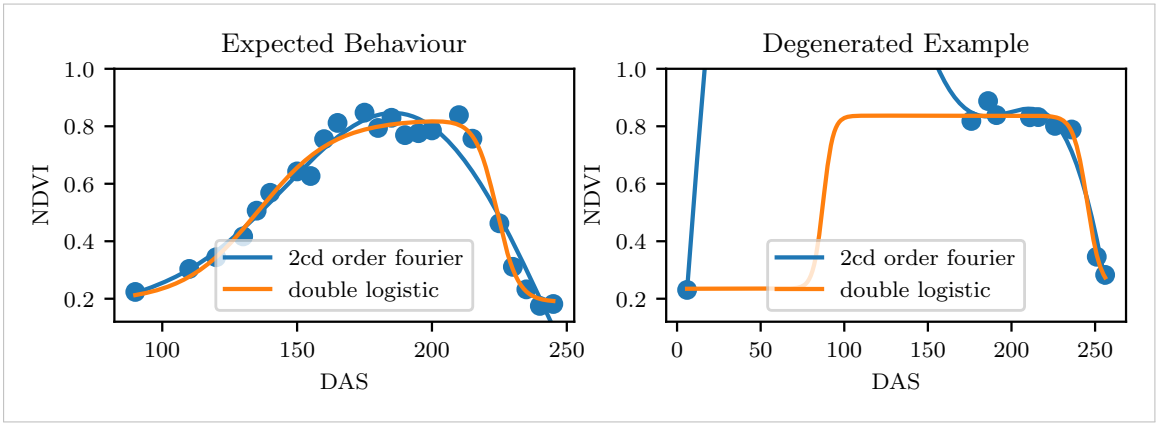


Figure 1.1: Here we observe the nice fitting possibilitys of the two parametric methods but notice also some misbehaviour

### 1.1.6 B-splines

from Lyche and Morken (Lyche and Morken)

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

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

**Smoothing:** We can relax the constrain that we have to perfectly interpolate. Thus we use the minimum number of knots[7] such that: $\sum_{i=1}^{n}(w(y_i - \hat{y}_i))^2 \leq s$

| Pros | Cons |
|---|---|
| — can be assigned degrees of freedom<br>— extendable to "smooth" version<br>— performs also well if points are not equidistant | — smoothing process does not translate well to a interpretation (unlike smoothing splines)<br>— choice of smoothing parameter $s$ |

[7] SciPy uses FITPACK and DFITPACK, the documentation suggests that smoothness is achieved by reducing the number knots used

### 1.1.7 Natural Smoothing Splines

Let $\mathcal{F}$ be the Sobolev space (the space of functions of which the second derivative is integrable). Then the unique[8] minimizer

$$\hat{m} := \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{n} (Y_i - f(x_i))^2 + \lambda \int f''(x)^2 dx$$

is a natural[9] cubic spline (i.e. a piecewise cubic polynomial function). The objective function has an intuitive meaning, as to avoid lateral acceleration it is desirable to move the steering wheel as little as possible, when driving a car.

| Pros | Cons |
|---|---|
| — can be assigned degrees of freedom (trace of the hat-matrix) | — choose $\lambda$ |
| — efficient estimation (closed form solution) | |
| — intuitive penalty (we don't want the function to be too "wobbly" — change slopes) | |
| — performs also well if points are not equidistant | |
| — fixes the Runge's phenomenon (fluctuation of high degree polynomial interpolation) | |

### 1.1.8 Whittaker Smoother

| Pros | Cons |
|---|---|
| — 1 | — 1 |
| — 2 | — 2 |

### 1.1.9 Kriging

Kriging was developed in geostatistics to deal with autocorrelation of the response variable at nearby points. By applying the notion that two spectral indices which are (timewise) close should also take similar values we justify the application of Kriging. In the end we would like to fit a smooth Gaussian process to the data. For this subsection we will follow Diggle and Ribeiro (dig).

#### Definitions and Assumptions

A *Gaussian Process* $\{S(t) : t \in \mathbb{R}\}$ is a stochastic process if $(S(t_1), \ldots, S(t_k))$ has a multivariate Gaussian distribution for every collection of times $t_1, \ldots, t_k$. $S$ can be fully characterized by the mean $\mu(t) := E[S(t)]$ and its covariance function $\gamma(t, t') = \mathrm{Cov}(S(t), S(t'))$

[8] Strictly speaking it is only unique for $\lambda > 0$

[9] It is called natural since it is affine outside the data range ($\forall x \notin [x_1, x_n] : \hat{m}''(x) = 0$)

Assumption: We will assume the Gaussian process to be stationary. That is for $\mu(t)$ to be constant in $t$ and $\gamma(t, t')$ to depend only on $h = t - t'$. Thus, we will write in the following only $\gamma(h)$.[10]

We also define the variogram of a Gaussian process as

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

And decide to use a gaussian Variogram defined by

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

where $h$ is the distance, $n$ is the nugget, $r$ is the range and $p$ is the partial sill visuilized in figure 1.2.[11]

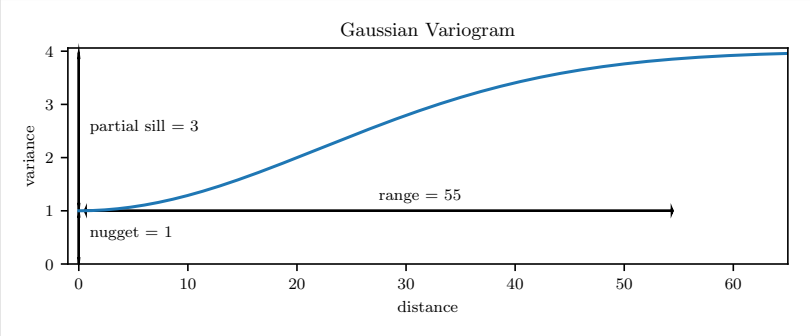


Figure 1.2: Gaussian Variogram with nugget=1, partial sill=3, range=55

| Pros | Cons |
|---|---|
| — It is a well-studied method | — Regression to the mean |
| — Parameters have an intuitive meaning | — Assumption of constant mean and constant variance is clearly violated. Thus, the NVI is not a stationary process. |
| — Flexible covariance structure | — Skewness of errors is not taken into account. |

### 1.1.10 Other Methods to study:

From inroduction of Chen, Jönsson, Tamura, Gu, Matsushita, and Eklundh (Chen et al.): (1) threshold- based methods, such as the best index slope extraction algorithm (BISE) (Viovy et al., 1992); (2) Fourier-based fitting methods (Cihlar, 1996; Roerink et al., 2000; Sellers et al., 1994); and (3) asymmetric function fitting methods such as the asymmetric Gaussian function fitting approach (Jonsson Eklundh, 2002) and the weighted least-squares linear regression approach (Swets et al., 1999).

[10] 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.

[11] Strictrly speaking we use a scaled version of the variogram. Thus only the ratio of $p/n$ matters.

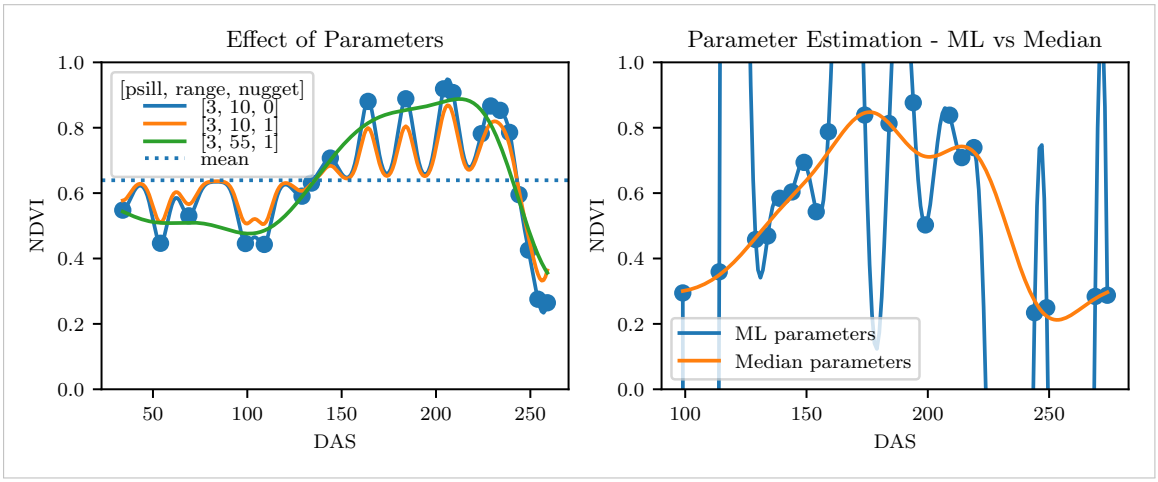


Figure 1.3: On the left we see how the interpolation change if we increase the nugget and the range parameter. On the right we compare two kriging interpolations where one takes parameters by numerically maximizing the (which results in a very small nugget) and the other takes the median of many such numerical optimizations.

# Bibliography


Gaussian models for geostatistical data. In P. J. Diggle and P. J. Ribeiro (Eds.), *Model-Based Geostatistics*, pp. 46–78. Springer.

Beck, P. S. A., C. Atzberger, K. A. Høgda, B. Johansen, and A. K. Skidmore. Improved monitoring of vegetation dynamics at very high latitudes: A new method using MODIS NDVI. *100*(3), 321–334.

Cao, R., Y. Chen, M. Shen, J. Chen, J. Zhou, C. Wang, and W. Yang. A simple method to improve the quality of NDVI time-series data by integrating spatiotemporal information with the Savitzky-Golay filter. *217*, 244–257.

Cappellari, M., R. M. McDermid, K. Alatalo, L. Blitz, M. Bois, F. Bournaud, M. Bureau, A. F. Crocker, R. L. Davies, T. A. Davis, P. T. de Zeeuw, P.-A. Duc, E. Emsellem, S. Khochfar, D. Krajnović, H. Kuntschner, R. Morganti, T. Naab, T. Oosterloo, M. Sarzi, N. Scott, P. Serra, A.-M. Weijmans, and L. M. Young. The ATLAS3D project - XX. Mass-size and mass-sigma distributions of early-type galaxies: Bulge fraction drives kinematics, mass-to-light ratio, molecular gas fraction and stellar initial mass function. *432*, 1862–1893.

Chen, J., P. Jönsson, M. Tamura, Z. Gu, B. Matsushita, and L. Eklundh. A simple method for reconstructing a high-quality NDVI time-series data set based on the Savitzky-Golay filter. *91*, 332–344.

Cleveland, W. S. Robust Locally Weighted Regression and Smoothing Scatterplots. *74*(368), 829–836.

Lyche, T. and K. Morken. Spline Methods.

Savitzky, A. and M. J. E. Golay. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *36*(8), 1627–1639.

Schafer, R. W. What Is a Savitzky-Golay Filter? [Lecture Notes]. *28*(4), 111–117.

# Appendix A

# Hi Mom