

MOTH AND BIRCH

DETECTION OF MOTH-INDUCED DEFOLIATION IN
BIRCH FOREST, USING REMOTE SENSING

VALENTINA PIVOTTI

Master's thesis
2017:E43



LUND UNIVERSITY

Faculty of Science
Centre for Mathematical Sciences
Mathematical Statistics

Abstract

The goal of this thesis is to build a near-real time defoliation detector that could be used, early in the spring, to find out which areas of a birch forest are being affected by an insect outbreak. The importance of a reliable detection at the beginning of the spring lies on the possibility for an early intervention.

The data on which the study is based is the Normalized Difference Vegetation Index (NDVI), which is calculated from the 8-days interval measurements, obtained from the MODerate resolution Imaging Spectroradiometer (MODIS). Further available information that is included in the analysis are the forest fraction and the altitude of each pixel.

The first step of the analysis is fitting a function to the measurements of each pixel whose parameters could capture the important aspects of the changes in NDVI values. Among them, the final NDVI value that is reached during the summer is of particular interest since an abnormally low value could be the indication of an insect infestation.

Different assumptions are made on the error distribution. The first ones, more simplistic, do not manage to counteract the noise. A more complex error structure is thus taken into account, leading to a better estimate which is then used to build the estimator.

The idea behind the detector is to identify those pixels for which the NDVI does not reach its high late spring/summer values fast enough, with respect to other pixels and previous years. It is known that the two years, among those available in the dataset, that have suffered a moth outbreak are 2004 and 2013. Hence, the estimation of the fitting function is run for 2000-2003, the detector is tried on 2004 and finally tested on 2013, since for this year a few locations of the outbreak were known. The discrepancy between field data and the results generated by the detector suggests further adjustments that would improve the capacity to detect moth infestation.

Contents

1	Introduction	1
1.1	Case Study	2
1.1.1	The National Park	2
1.1.2	<i>Epirrita Autumnata</i>	2
1.2	Remote Sensing	3
1.2.1	MODIS	3
1.2.2	NDVI	4
2	Data	5
2.1	Raw data	5
2.2	Data pre-processing	7
3	Models	9
3.1	First modeling attempts	9
3.1.1	Theory	10
3.1.2	Results and Observations	12
3.1.3	Parameters distribution	15
3.2	Advanced models	17
3.2.1	MCMC	17
3.2.2	Quantile regression on GH	20
4	Near real-time detector	24
4.1	Procedure	24
4.1.1	Detector 1	26
4.1.2	Detector 2	28
4.1.3	Detector 3	30
4.1.4	Comparing detectors	31
4.2	Check detector on 2013 data	32
4.2.1	Actual outbreak	32
5	Conclusions	34
5.1	Models	34
5.2	2013 Detector	34

Chapter 1

Introduction

Healthy forests are of great importance for absorbing atmospheric CO₂, thus keeping its concentration under control and limiting its effects on global warming. Moreover, in the Scandinavian peninsula, forest related industries such as wood and paper production, represent a preponderant source of income: the total export of forestry products for the year 2017, between January and March, was of more than 33.5 billion Swedish krona [1], positioning the forest industry at the third place for export, among other commodity groups. This economic relevance adds further importance to forest health preservation.

Pests infestation and the subsequent induced defoliation are one of the biggest threat to their health since they could lead to long standing damages in those areas that have been affected. It is therefore very important to constantly monitor them [2]. Frequent information about forest conditions over large areas is thus needed because it allows for a broader understanding of pests dynamics and, when needed, repentine counter-measures [3]. Early interventions could be they key to save large portions of forests before the outbreak reaches its peak. A very valuable tool in this kind of defoliation monitoring are Remote Sensing techniques which can provide measurements collected on a regular basis over wide areas.

The aim of this project is to build a near real-time detector of defoliation in the specific case study of the birch forest of the National Park of Abisko, in Northern Sweden. In particular, the goal is to identify the occurrence of abnormal behaviors during the spring bud burst. Alterations from the norm indicate, in fact, defoliation which could be the sign of local outbreaks of the autumnal moth *Epirrita Autumnata*. This investigation utilizes satellite based measurements of MODIS¹ derived NDVI² of the National Park recorded between 2000 and 2014.

¹MODerate resolution Imaging Spectroradiometer

²Normalized Difference Vegetation Index



[4]

1.1 Case Study

1.1.1 The National Park

The National Park of Abisko consists of 77 km² of birch forest and alpine flora, on the shores of the lake Torneträsk. It is located in the municipality of Kiruna, 200 km North of the Arctic Polar Circle [5]. Founded in 1909, the park is a growing touristic destination and an important center for scientific research. The Abisko Scientific Research Station, built in 1913, has been carrying out leading studies and collecting records of environmentally relevant measurements in the region, since its foundation [6]. Scientific research in the polar area is of high importance on its own, but particularly because of the strong impact that climate change has at these extreme latitudes.

1.1.2 *Epirrita Autumnata*

Epirrita Autumnata is a geometer moth, commonly found in the birch forests of the Scandinavian peninsula, together with the winter moth (*Operophtera brumata*). The name geometer refers to the ways larvae move forward. They proceed extending the full length of their body, before deciding where to head to, and this movement looks like they measure (meter) the Earth (geo). The moth's life cycle consists of four phases: egg, larva, pupa and adult. In September, adult moths lay their eggs that rest during the cold season until they hatch at the following bud burst, giving start to the larval stage. This is the phase when the insect feeds on birch leaves. From the end

of June, the larvae move to the next stage and pupate on the forest ground until the end of August when the adults mate and the cycle restarts [7]. Apart from its diet on birch leaves, another reason why the life cycle of this insect has been widely studied, is because its variations have been proven to be related to rising temperatures. In particular, mild winters favor eggs survival, allowing the following moth generation to expand its territory to areas previously adverse for the insect [7].

1.2 Remote Sensing

Remote sensing methods have been developing since the launch of the first Landsat satellite in 1972 [8]. These techniques come as a great help in a study that seeks to monitor large areas like the Abisko National Park. In fact, observations in situ would be highly expensive and they could never cover the whole surface with the high time resolution that MODIS allows for. In addition, the measurements provided from MODIS go beyond human sight. The light spectrum that is recorded is wider than visible light and allows for a better understanding of tree health. Furthermore, abundant literature describes the successful use of remote sensing for defoliation detection which is the aim of this project. In a review from 2013 [2], it is concluded that MODIS is the best sensor, so far, for early detection of defoliation thanks to its high time resolution. This high frequency is paid by having low spatial resolution. In the review, however, it is stated that this trade-off is an acceptable compromise in an environment like the Scandinavian peninsula where the forest is almost uninterrupted.

1.2.1 MODIS

This acronym stands for MODerate resolution Imaging Spectroradiometer and it is the name of the instrument used to collect images of the Earth surface at different spectral bands. MODIS is located on two different satellites, Terra and Aqua, which collect images of our planet with a time interval between 1 and 2 days, depending on the latitude [9]. Both satellites travel between poles, in opposite directions, passing the Equator at different moments, in order to get high coverage of Earth surface[10]. The data available from MODIS can be divided into three categories: spectral, temporal and angular. In other words, remotely sensed images are collected at a wide light spectrum and the time resolution of the observations is available together with the angle between the satellite and the Earth surface in the moment the image is recorded. Among all measurements collected by MODIS, this project uses MOD09, which is the surface spectral reflectance in the visible and near infra-red light. This data is then transformed into NDVI, the details are described in Section 1.2.2 .

At latitudes as high as the one in Abisko, MODIS collects images daily. However, the data used in this analysis consists of one observation every 8 days. For each time interval the measurement kept is the one presenting the highest NDVI value. Keeping the 8-day maximum is a way for discarding observations corrupted by noise. Since the interference, mainly due to cloud, atmospheric absorption and background color variations, usually has a negative bias, considering only the highest NDVI value within each 8-days interval is an effective way of reducing noise [11].

1.2.2 NDVI

The acronym NDVI stands for Normalized Difference Vegetation Index, a well established toll used to measure the presence of healthy green vegetation [12]. The formula to calculate it is the following ratio:

$$\text{NDVI} = \frac{\text{NIR} - \text{R}}{\text{NIR} + \text{R}} \quad (1.1)$$

where NIR stands for the measured reflectancies of Near Infra-Red light and R for reflectancies in Red [13]. The index uses these two light channels because green leaves absorb light in the visible red band (0.5 - 0.7 μm) and reflect in the near infra-red wavelength interval (0.7 - 1.1 μm). For this reason, a ratio calculating the relation between the two channels helps identify live green mass. However, a simple ratio NIR/R between the two channels would create an instable index, that could be highly affected by measurement errors and noise, while the normalization in the formula guarantees a more stable detector [13]. NDVI is a widely used indicator for green mass, but there are also other vegetation indexes, each one developed in response to a specific research need or problem. As an example, the Transformed Vegetation Index (TVI) is a variation of the NDVI, $\text{TVI} = \sqrt{\text{NDVI} + 0.5}$, that was created in order to avoid negative values [12].

Chapter 2

Data

2.1 Raw data

As mentioned before, the raw data used in this project is the 8-days filtered NDVI calculated from MOD09 on a surface of roughly 350 km^2 around the Abisko national park, which corresponds to 50×110 pixels, each $250\text{m} \times 250\text{m}$ wide. As far as the filtering, it is important to point out that, for each 8-days interval, the day with the highest value could differ between pixels, thus making the time series of "accepted days" different for each pixel. This happens because, for example, disturbances such as clouds can have a local, very limited interference and affect only a small amount of pixels at a time. In Figure 2.1 an example of raw data from the year 2000 is plotted for the pixel in position $(25, 55)$, which is central in the grid. The time interval matching the beginning of the green season is marked in red.

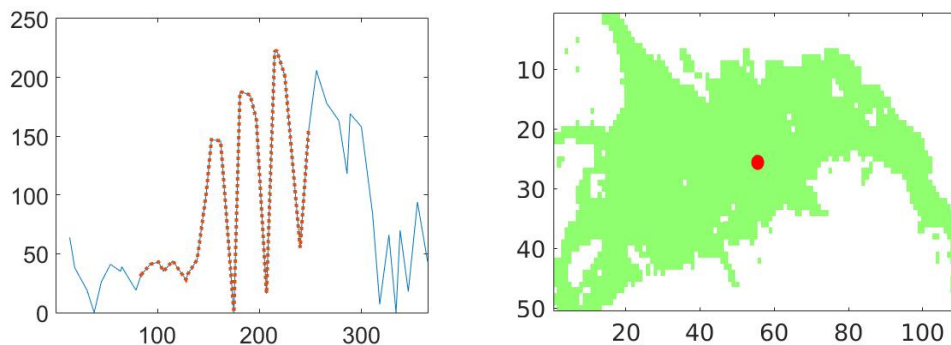


Figure 2.1: On the left, example of raw data, the red dashed line represents the time from spring bud burst to fully developed leaves in the summer. The project's time interval of interest. On the right, the position of this specific pixel within the park is shown.

Additional information relevant for the study is the quality indicator of

each measurement, equal to either 0 (poor), 1 or 2 (good). In particular, through the whole dataset, between 2000 and 2014, only 30% observations presented positive quality (13% equal to 2, 17% equal to 1). This percentage can appear low if one only thinks about occasional bad weather condition, but it is important to keep in mind that during the winter season the snow covers the land, causing the reflectance image to be corrupted for the whole duration of the season. As a result, the quality of the NDVI is consistently null for most part of the year. This fact can be seen in Figure 2.2 where the quality is null for all observations until the end of April (8-days interval ~ 30). During the summer the observations improve, reaching the peak of high quality concentration around July-August (8-days interval ~ 210), and subsequently worsen during the Autumn.

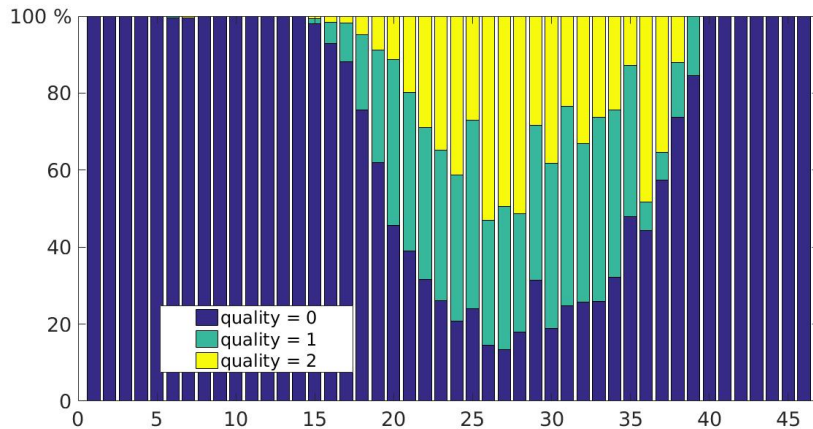


Figure 2.2: Percentage of different qualities of the observations for each 8-days interval throughout the years 2000-2014.

Furthermore, the altitude value for each pixel of interest is provided, together with its forest fraction. High altitude and low forest fraction occur simultaneously in an environment like the Abisko National Park where the tree line coincide with rising ground. Therefore, in the dataset available for the project, all pixels having altitude higher than 1007 m have been removed. In Figure 2.3, both altitude and forest fraction are presented, the pixels outside the area of interest are colored in white. One can notice the gradient of the altitude increasing towards the white areas in the South and West which are mountains, while it decreases towards the white region in the North-East, which is the lake Torneträsk. It is interesting to note how the forest fraction decreases both when the altitude rises and when approaching the shores of the lake.

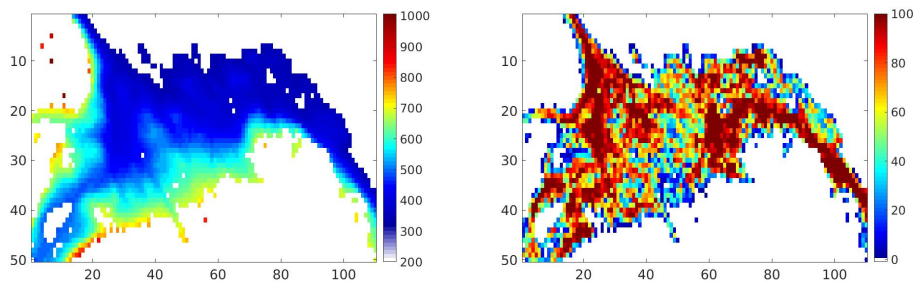


Figure 2.3: On the left the altitude, on the right the forest fraction

Finally, the daily average temperature is included. This data could, in fact, provide insights about the beginning of the season since at this latitude spring depends primarily on the occurrence of milder temperatures. This can be seen in Figure 2.4 where the NDVI grows together with the daily temperature and decreases once the autumn temperature starts dropping.

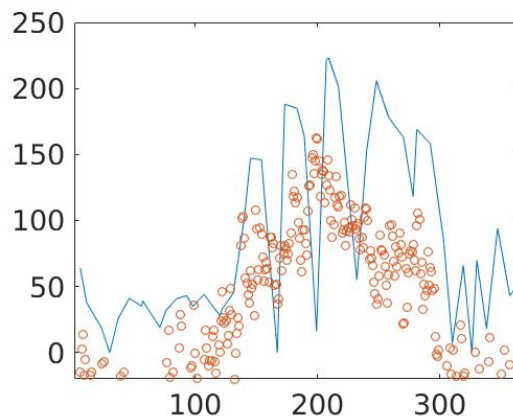


Figure 2.4: Raw data against daily temperature (the latter is multiplied by a factor of 10 for illustrative purposes).

2.2 Data pre-processing

The first adjustment of the data was to match each observation with its yearly day which guarantees a more thorough description of the seasonal changes, especially considering that the main focus of the project is to capture the steep growth during the spring which happens within a few weeks.

Secondly, altitude and forest fraction needed to be taken into account. As mentioned before, all pixels for which altitude was higher than 1007m were discarded. From the 5500 original pixels, the dataset was reduced to 2541 pixels.

Finally, several approaches were tried out in order to utilize the quality information and thus reduce the influence of those measurements that are corrupted by noise. One attempt, presented here as an example, was to remove the measurements with quality equal to 0. However, as said before, only 30% of the observations have a positive quality and so the removal of null quality observations reduced the dataset too drastically. In particular, the winter season lost almost all observations, making it impossible to detect changes in vegetation caused by the start of the growing season during early spring. Among all adjustments that were tried, none brought an improvement so significant that could justify such a strong intervention on the data.

Chapter 3

Models

3.1 First modeling attempts

The three methods described in the following section attempt to fit a function to the observations within the time interval March-September, when the NDVI increases and then stabilizes. The function chosen to be fitted, is the following:

$$f(\vec{t}_i, \Psi_i) = k_i + \frac{a_i}{1 + e^{-b_i(\vec{t}_i - t_{0i})}},$$

where $\vec{t}_i = (t_{i1}, t_{i2}, \dots, t_{in})$ is the time vector for the i -th pixel and $\Psi_i = (a_i, b_i, t_{0i}, k_i)$ are pixel-specific parameters. The choice of this function depends on the fact that its parameters capture all the interesting features of the seasonal change in each time series. In particular, the parameter k identifies the basis value of NDVI during the winter season while a represents the value that is reached after the blooming interval and maintained throughout the summer. Furthermore, the parameter b captures the speed at which the vegetation index grows and t_0 gives an indication of when the bud burst happens. In fact, when the time point is equal to t_0 , the function is equal to $k + a/2$, meaning that at that time point the growth of NDVI is half way. It is interesting to focus on the impact the b parameter has on the shape of the fitting function. On the left side of Figure 3.1, different functions are plotted, they all have $a = 160$, $k = 40$ and $t_0 = 150$, while b varies from 0.05 up to 0.8. It can be seen that most variation happens for values between 0.03 and 0.5, for smaller values the function approaches a straight line, while for higher values the variation between functions is minimal. The function $f(\vec{t}_i, \Psi_i)$ has symmetrical slope, therefore, at a later stage, a fifth parameter ν_i that allows for an asymmetric slope is included as follows:

$$g(\vec{t}_i, \Theta_i) = k_i + \frac{a_i}{(1 + \nu_i e^{-b_i(\vec{t}_i - t_{0i})})^{1/\nu_i}}.$$

On the right side of Figure 3.1, one can see the impact of ν (the other parameters are the same as on the left side with $b = 0.1$) on the function $g(\vec{t}_i, \Theta_i)$.

It is interesting to point out that $g(\vec{t}_i, a_i, b_i, t_{0i}, k_i, \nu_i = 1) = f(\vec{t}_i, \Psi_i)$.

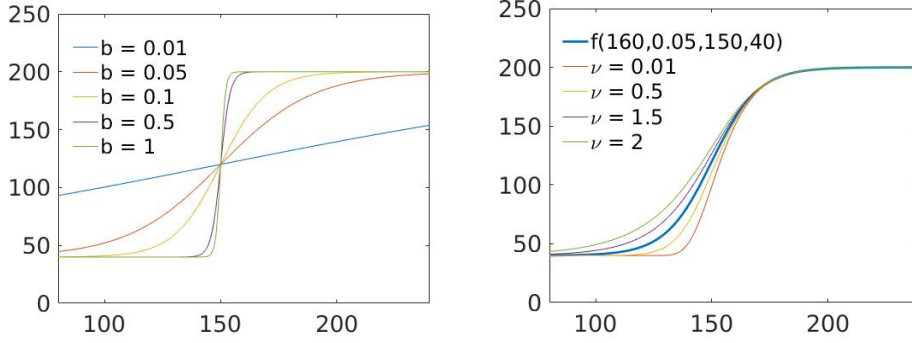


Figure 3.1: On the left, the function $f(\vec{t}_i, \Psi_i)$ for different values of b . On the right, the function $g(\vec{t}_i, \Theta_i)$ for different values of ν .

All methods are tried several times on different years, pixels and time intervals. The interval that appears to be the best choice is $[10, 30]$. It provides enough observations allowing the function to capture both the low values at the end of winter and the high late spring/early summer ones. Both extremes of the interval indicate the number of the 8-day slot, meaning that the interval covers the period between the end of March (\sim day 80) and the end of August (\sim day 240). It is the interval marked in red in Figure 2.1.

Even if all pixels are included in the study, the results of only four of them are plotted. In order to illustrate the whole park, four pixels are picked spread throughout the grid, at different altitudes and forest fractions.

pixel	altitude	forest fraction
29 – 8	767	9
33 – 42	550	47
25 – 55	470	77
27 – 102	345	21

3.1.1 Theory

Non-linear Regression At first the parameters Ψ_i are estimated using the Least Squares principle

$$\hat{\Psi}_i = \arg \min_{\Psi_i} \sum_{j=1}^n (y(t_{ij}) - f(t_{ij}, \Psi_i))^2,$$

where the parameter j indicates the time points of each vector \vec{t}_i and i indicated the site. This is the first, simplistic approach. In order for these estimated $\hat{\Psi}_i$ to be MLE, the assumption is that the residuals are normally distributed and homoscedastic.

Secondly, the estimation is repeated using Least Absolute Deviation. The reason is to use a more robust estimating procedure, being less affected by outliers, since the data used for this analysis is very noisy. In this case, the assumption under which $\hat{\Psi}_i$ is MLE is for the residuals to be from a Laplace distribution.

$$\hat{\Psi}_i = \arg \min_{\Psi_i} \sum_{j=1}^n |y(t_{ij}) - f(t_{ij}, \Psi_i)|.$$

The results of both estimated $\hat{\Psi}_i$ and their fitting functions can be seen in the summarizing Figure 3.3.

The densities of the residuals are plotted in Figure 3.2. The one on the left is from the LS residuals and it should be Gaussian. It appears though that it is too sharp to belong to a normal distribution and its tails are too heavy. Similarly, the density on the right side, whose sharpness could correspond to a Laplace distribution, appears to have an asymmetric curve around the value 50 that the 2-sided exponential distribution could not capture.

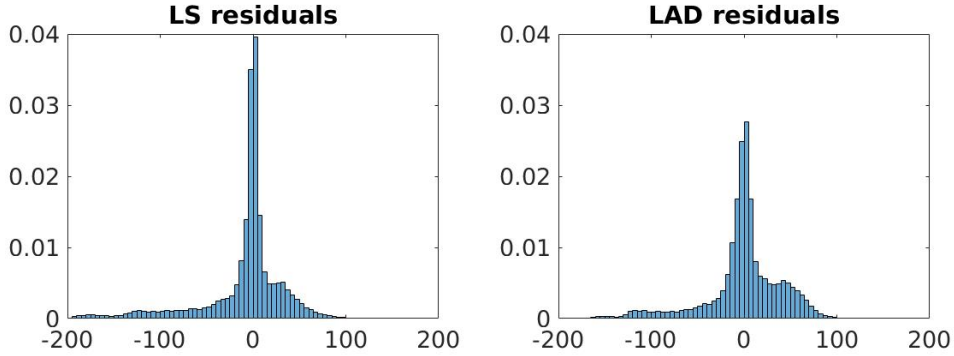


Figure 3.2: Empirical densities of the residuals from the first fitting methods.

Re-weighted Regression The first two methods are not a good fit for the data because of violation of assumptions. A third approach is then tried out where the heteroscedasticity of the residuals is included in the estimating process. The variance of the residuals can be written as:

$$\epsilon(t_{ij}) = y(t_{ij}) - f(t_{ij}, \Psi_i) \quad \text{Var}(\epsilon(t_{ij})) = \sigma^2(a_{i,1}^2, \dots, a_{i,n}^2) \quad \forall i.$$

This means that a model where all terms are multiplied by their weight $a_{i,j}^{-1}$ for each pixel i and time point j would respect the homoscedastic assumption [14]. The residuals would then be:

$$\epsilon(t_{ij})^* = a_i^{-1}(y(t_{ij}) - f(t_{ij}, \Psi_i)) \quad \text{Var}(\epsilon(t_{ij})^*) = \sigma^2(1, \dots, 1) \quad \forall i.$$

The weights are chosen to be estimated as $\frac{1}{|y(t_{ij}) - f(t_{ij}, \Psi_i)|}$, meaning that those observations whose distance from the fitting function is higher, weight less in the least square minimization. The iteration is run until the estimates find an equilibrium:

$$\hat{\Psi}_i^{(k+1)} = \arg \min_{\Psi_i} \sum_{j=1}^n (y(t_{ij}) - f(t_{ij}, \Psi_i))^2 \frac{1}{|y(t_{ij}) - f(t_{ij}, \hat{\Psi}_i^k)|}$$

In the final Figure 3.3 one can see the results compared to the previous methods.

Quantile Regression The last among these first estimators is the quantile regression. It is a more robust estimator which, depending on the choice of τ , estimates the $\hat{\Psi}_i$ such that $f(\vec{t}_i, \hat{\Psi}_i)$ is the approximation of the τ th quantile of $y(\vec{t}_i)$ [15]. The function to minimize for each pixel is the following sum along the time vector \vec{t}_i :

$$\hat{\Psi}_i = \arg \min_{\Psi_i} (F_q(a_i, b_i, t_{0i}, k_i)) = \arg \min_{\Psi_i} \sum_{j=1}^n \rho_\tau(\epsilon(t_{ij})), \text{ where}$$

$$\epsilon(t_{ij}) = y(t_{ij}) - f(t_{ij}, \hat{\Psi}) \text{ and } \rho_\tau(u) = u(\tau - I(u < 0)).$$

As described in previous sections, the bias affecting the observations is mostly negative; therefore, the values of τ that are tried out are all > 0.5 . On the right side of Figure 3.3, the optimizations for $\tau = 0.65$ and 0.75 are plotted.

A further attempt is tried out in order to reach a higher stability and reduce the influence of occasional bad measurements: the quantile regression applied on groups of pixels. In particular, for the estimation of each Ψ_i , the 25 pixels around the i th one are included in the minimization. The estimation is then:

$$\hat{\Psi}_i = \arg \min_{\Psi_i} \sum_i^{25} (F_q(a_i, b_i, t_{0i}, k_i)) = \hat{\Psi}_i = \arg \min_{\Psi_i} \sum_{i=1}^{25} \sum_{j=1}^n \rho_\tau(\epsilon(t_{ij})),$$

where the neighborhood consisted of a 5x5 square around the pixel for which the estimation was done. The resulting functions are also plotted on the right of Figure 3.3.

3.1.2 Results and Observations

First of all, the estimations of Ψ_i done with LS and LAD do not hold because the residuals do not appear to belong to the needed distributions, as previously discussed. The plots are included anyways on the left side of the figure and it appears clear that both estimations are highly influenced by the

noise during the green season. It seems in fact that the functions try to find a "compromise" between corrupted and non-corrupted measurements and do not reach the high levels of summer peaks which are the main interest of the analysis.

Secondly, the re-weighted regression performs very well in some cases and very poorly in other; among the examples, one can see good fitting in the first two cases and poor fitting in the latter two.

On the right side, instead, one can immediately notice the improvement provided by the quantile regression: both methods capture very well the rise and the starting point of the season, and manage to reach the high peaks during the summer (higher when $\tau = 0.75$). There is, however, room for improvement. Corrupted measurements are still affecting the fitting functions too much and it appears that the function could need an asymmetric rise, since the second part of the function appears to be smoother than the first one. This can be seen in Figure 3.3, especially for pixels 33-42 and 25-55. The function $f(\vec{t}_i, \Psi_i)$ is in fact able to capture well either the beginning or the end of the slope, but not both. These improvements to the fitting model will be applied and discussed in later sections.

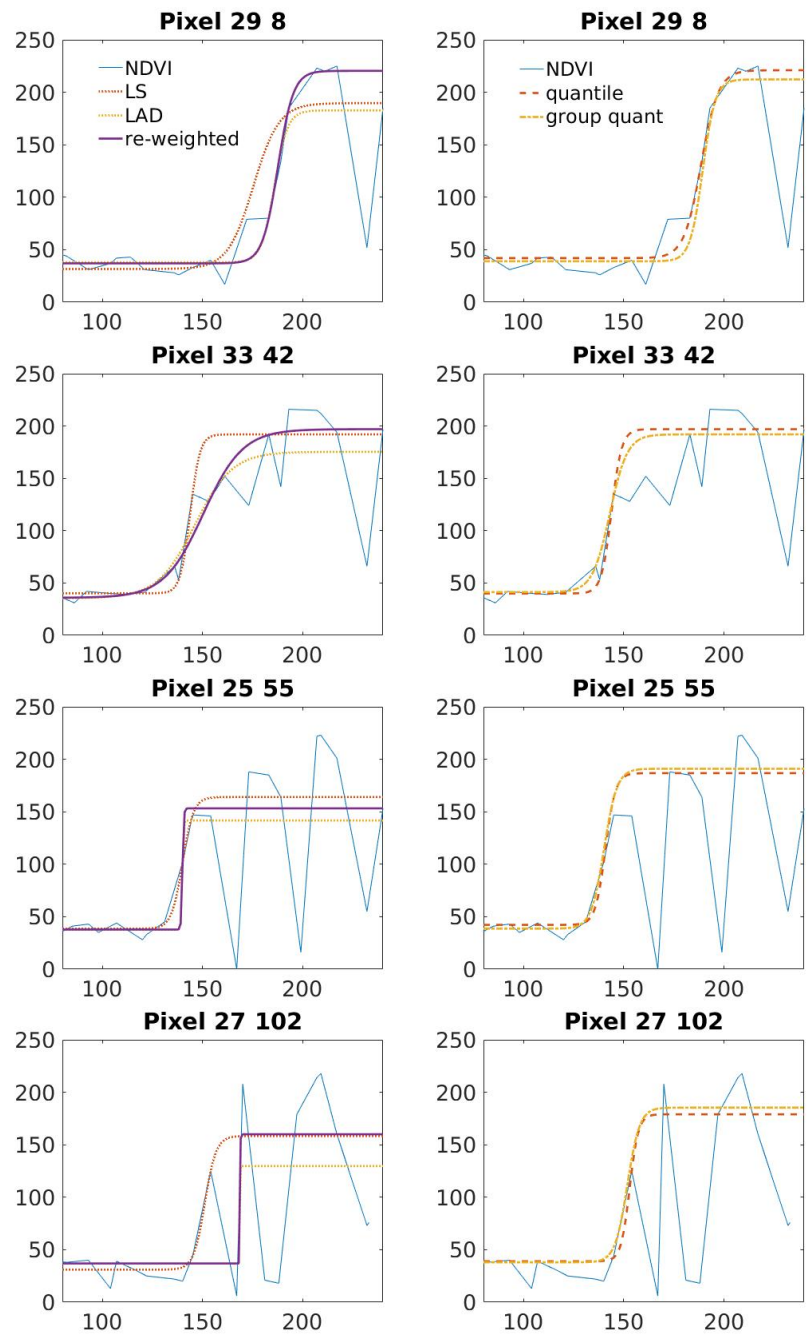


Figure 3.3: Fitting functions obtained, on the left side, from the minimization of the residuals' norm and the re-weighted regression; on the right side instead, the plots are a result of the different quantile regressions. The legend in the plots of the first pixel is valid for all four.

3.1.3 Parameters distribution

In order to gather further information, the spatial distributions of all parameters Ψ_i over the whole area of interest are studied. Since the quantile regression is the model that performed best so far, the parameters from both the single and the grouped-pixel quantile regression are plotted in Figure 3.5. In particular, the choice of $\tau = 0.65$ seemed the most reasonable compromise and is therefore used in these figures.

There are several relevant observations that can be made on the resulting plots. First, it is interesting to notice that, between the two methods, the parameters have comparable spatial distributions. The main difference is the higher smoothness in the case of grouped pixels, as it was expected. The only parameter that significantly changes between methods is b which is on average lower in the grouped pixels approach. One possible explanation may be that, when more pixels are taken into account, sudden rises are averaged. This behavior can be seen in Figure 3.4 where the first peak is ignored when the estimation is done on groups and the function has a totally different slope than in the regular case. This behaviour suggests that the option of grouped pixels could be of help in handling low quality, corrupted measurements. Furthermore, an interesting relation can be seen between the distribution of parameter a and forest fraction from Figure 2.3. At the same time there seems to be a possible positive correlation between t_0 and altitude, also in Figure 2.3, together with a negative correlation between b and altitude. Finally, k varies very little throughout the whole park surface. All these properties will be studied in later sections and some of them used in the detector.

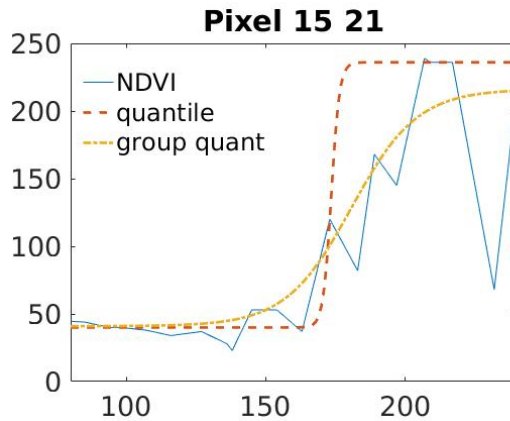


Figure 3.4: Quantile and grouped quantile regression with $\tau = 0.65$.

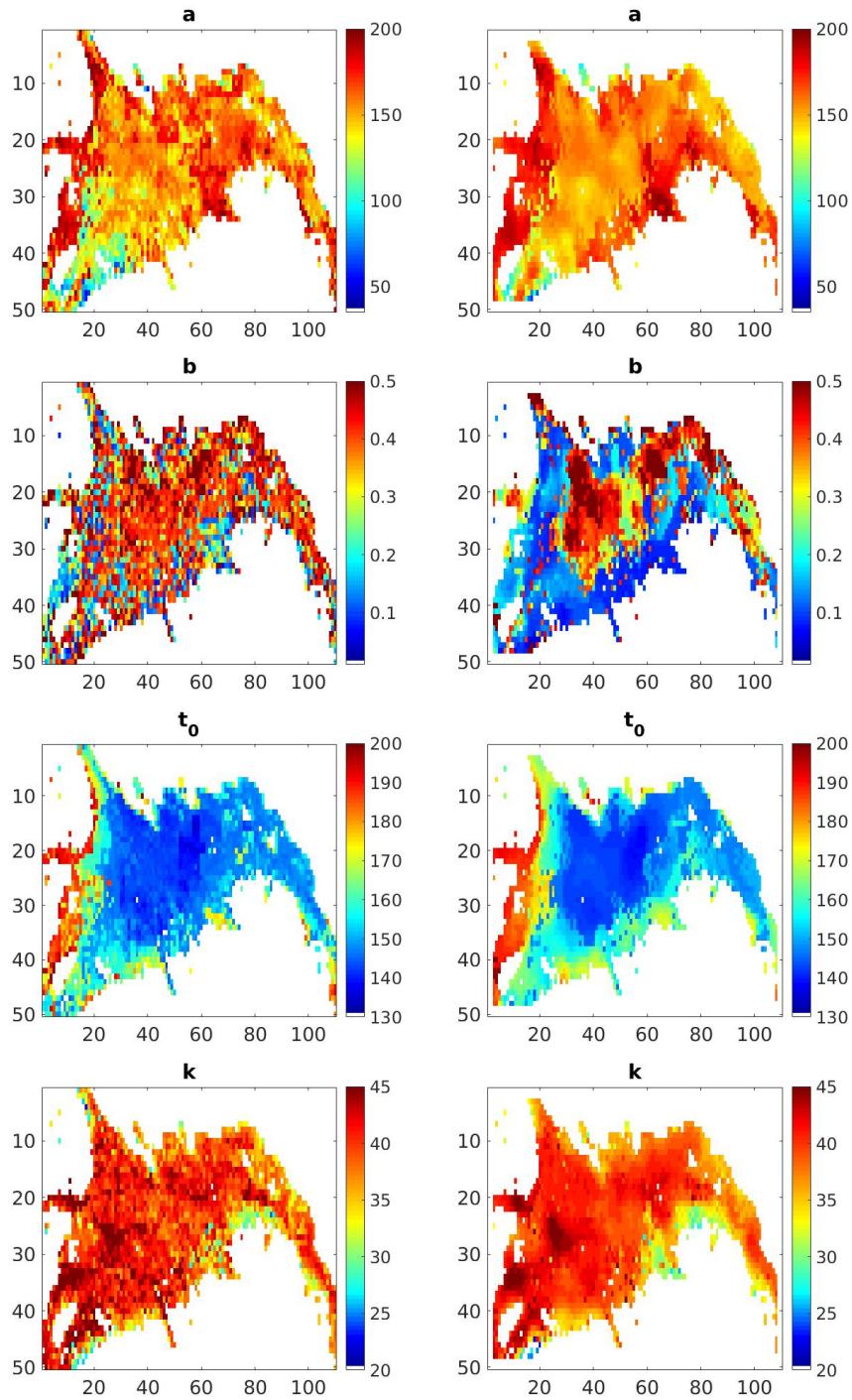


Figure 3.5: The spatial distribution of the parameters of the function fitted with quantile regression, for single pixels on the left side, for grouped pixels on the right.

3.2 Advanced models

All the methods described in the previous section are highly influenced by the noise. Throughout the figures it can be seen that the fitting function is either unable to capture the beginning of the season, or it is affected by low measurements occurring in the late spring/summer. Therefore more sophisticated tools need to be tested. In particular, the models that follow are different implementations of the same idea: the error structure need to be more complex in order to contrast the effect of noise.

The errors are then modelled according to the following Generalized Hyperbolic distribution [16]:

$$\begin{aligned} y_{ij} &= f(t_{ij}, \Psi_i) + \theta z_{ij} + e_{ij} \sqrt{\frac{z_{ij}}{\sigma}} \\ e_{ij} &\sim N(0, 1) \quad z_{ij} \sim GIG(p, a, b). \end{aligned} \quad (3.1)$$

The parameters of the GIG-distribution are set equal to: $p = -1/2$, $a = \alpha^2$ and $b = 1$ which is equal to say that the distribution of the terms z_{ij} is Inverse Gaussian $IG(\alpha^2, 1)$ with density equal to:

$$f(z|\alpha) = \frac{1}{2K_{-1/2}(|\alpha|)} \frac{1}{\sqrt{\alpha z^3 e^{(\alpha^2 z + 1/z)}}},$$

where $K_{-1/2}(|\alpha|)$ is the modified Bessel function of the second kind [17]. It also defines the complete error structure $\theta z_{ij} + e_{ij} \sqrt{\frac{z_{ij}}{\sigma}}$ as a general NIG type distribution. Furthermore, it is important to point out that, since the terms e_{ij} belong to a standard normal distribution, y_{ij} is conditionally gaussian, with variance equal to $\frac{z_{ij}}{\sigma}$. The values z_{ij} are therefore capturing the heteroscedasticity of the residuals.

On this error structure, two different approaches are tried out.

3.2.1 MCMC

Theory

All parameters appearing in (3.1), with the exception of the coefficients Ψ_i , are estimated using Gibbs sampling from their posterior distributions.

The parameter σ , whose prior is assumed to be $\Gamma(a_\sigma, b_\sigma)$, has posterior distribution:

$$\sigma | \dots \sim \Gamma\left(\frac{Nn}{2} + a_\sigma, \frac{1}{2} \sum_{j=1}^n \sum_{i=1}^N \frac{(y_{ij} - f(t_{ij}, \Psi_i) - \theta z_{ij})^2}{z_{ij}} + b_\sigma\right),$$

where n is the number of time points j and N is the number of pixels i .

Similarly, the parameter θ , having prior $N(0, 1/q)$, has posterior:

$$\theta \sim N\left(\frac{\sum_{j=1}^n \sum_{i=1}^N (y_{ij} - f(t_{ij}, \Psi_i))}{q + \sigma \sum_{j=1}^n \sum_{i=1}^N z_{ij}}, (q + \sigma \sum_{j=1}^n \sum_{i=1}^N z_{ij})^{-1}\right).$$

As far as the terms z_{ik} , their posterior belongs to the same GIG family as their prior, but with the following parameters:

$$z_{ij} | \dots \sim GIG(-1, \theta^2 \sigma + \alpha^2, (y_{ij} - f(t_{ij}, \Psi_i))^2 \sigma + 1) \quad \forall i, j.$$

Furthermore, the posterior of α depends on the error type which, in this case, is NIG. This means that the logarithm of the posterior distribution of α is equal to:

$$\log p(\alpha | \dots) = (a_\alpha - 1) \log(\alpha) - \frac{\alpha^2}{2} \sum_{j=1}^n \sum_{i=1}^N z_{ij} + \alpha(n - b_\alpha).$$

The logarithm is concave, and this allows the sampling of α to be done using the Adaptive rejection sampling (ARS) [18].

Finally, the parameters of interest Ψ_i are considered to have the following prior distribution:

$$\Psi_i \sim N(\mu_\Psi, \Sigma_\Psi), \text{ where } (\mu_\Psi, \Sigma_\Psi) \sim NIW(\mu_0, k_0, \Lambda_0, \nu_0).$$

This means that mean and variance of Ψ belong to a Normal Inverse Wishart, in particular

$$\mu_\Psi \sim N\left(\mu_0, \frac{1}{k_0} \Sigma_\Psi\right) \quad \Sigma_\Psi \sim IW(\nu_0, \Lambda_0).$$

Their conditional posterior is therefore equal to:

$$(\mu_\Psi, \Sigma_\Psi | \dots) \sim NIW(\mu_n, k_n, \Lambda_n, \nu_n)$$

where

$$\mu_n = \frac{k_0}{k_n} \mu_0 + \frac{n}{k_n} \bar{\Psi}_i, \quad k_n = k_0 + n, \quad \nu_n = \nu_0 + n$$

and

$$\Lambda_n = \Lambda_0 + \sum_i (\Psi_i - \bar{\Psi}_i)(\Psi_i - \bar{\Psi}_i)^T + \frac{k_0 n}{k_n} (\bar{\Psi}_i - \mu_0)(\bar{\Psi}_i - \mu_0)^T$$

The sampling of Ψ_i is done using MALA (Metropolis-adjusted Langevin algorithm) [19; 20]. This is a tool that samples following a Metropolis-Hastings algorithm, but its proposed values for the random walk are created, based on the distribution of Ψ_i . In particular the algorithm uses gradient

and Fisher information matrix of the negative logarithm of the posterior distribution.

$$\begin{aligned}
-\log p(\Psi_i | \dots) &= -\log p(y | \dots) - \log p(\Psi_i) \propto \\
&\frac{1}{2} \sum_{j=1}^n (y_{ij} - f(t_{ij}, \Psi_i) - \theta z_{ij})^2 \frac{\sigma}{z_{ij}} - \log p(\Psi_i) \propto \\
&\frac{1}{2} \left(\sum_{j=1}^n (y_{ij} - f(t_{ij}, \Psi_i) - \theta z_{ij})^2 + (\Psi_i - \mu_\Psi)^T \Sigma_\Psi^{-1} (\Psi_i - \mu_\Psi) \right).
\end{aligned} \tag{3.2}$$

Implementation The actual implementation of the MCMC includes two steps. First of all, prior values for Ψ_i are estimated as:

$$\hat{\Psi}_i = \arg \min_{\Psi_i} (-\log p(\Psi_i | \dots)).$$

The function that runs the minimization is also given the gradient as a further input, in order to obtain better results. Both log-likelihood and gradient are from the same posterior distribution used for the MALA sampling. The first guess for Ψ_i provided to the function is the median of the $(\hat{\Psi}_i)_i$ estimated with the quantile regression of the previous section. In order to further simplify this preliminary estimation, θ is set equal to 0 and $z_{ij} = 1 \quad \forall i, j$, thus returning to a Gaussian model for the errors.

After these first steps, the actual MCMC iterations are started. Within the cycle, the MALA sampling is run first, based on those prior predictions for Ψ_i . In sequence, the remaining passages of the MCMC are run, sampling all other parameters from their posteriors. The whole MCMC estimation is run several times. According to the results, hyperparameters are tuned, particularly those for σ and θ . The results plotted here show the equilibrium that the estimates reached after roughly 1000 iterations.

The final fitting is plotted in Figure 3.6

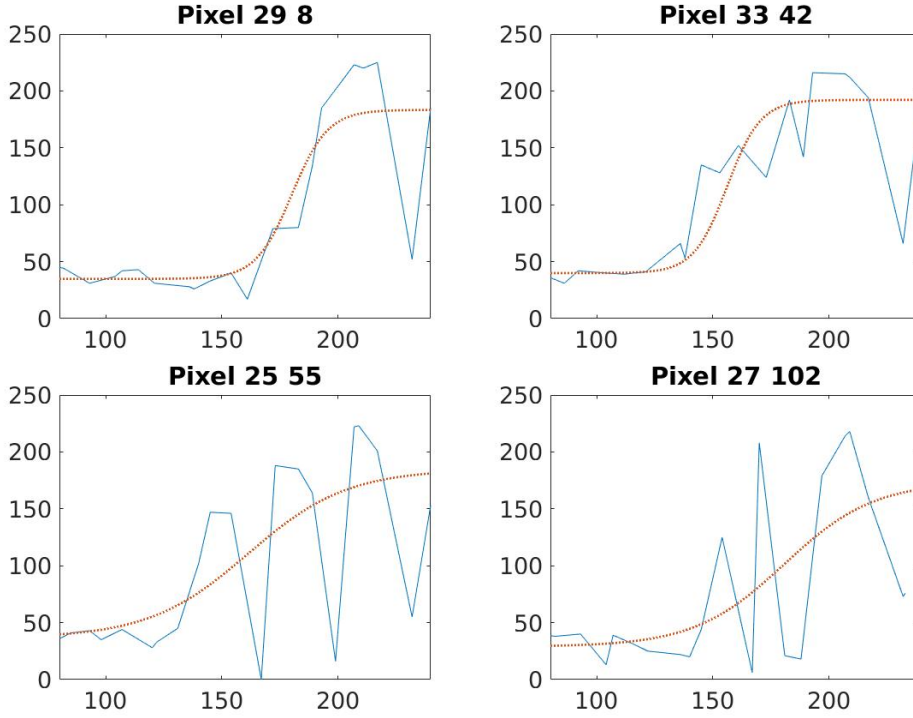


Figure 3.6: Fitting functions obtained from MCMC.

Results and Observations

It can be noticed, in Figure 3.6, how the Markov Chain Monte Carlo allows for a smoother fitting of the observations. All the peaks are interpolated and there is no steep growth at bud burst, as it was happening with previous models. However, the fitting function is too low, compared to the actual values of NDVI: the influence of corrupted measurements is still too high.

3.2.2 Quantile regression on GH

Theory

This second method is an original approach to the task: it seeks to combine the Generalized Hyperbolic error structure defined in (3.1) with the quantile regression. The result is as follows:

$$\hat{\Psi}_i = \arg \min_{\Psi_i} \sum_i^{25} F_q(\Psi_i) = \arg \min_{\Psi_i} \sum_i^{25} \sum_{j=1}^n \rho_\tau(\epsilon(t_{ij})), \quad (3.3)$$

$$\text{where } \epsilon(t_{ij}) = \frac{y_{ij} - f(t_{ij}, \Psi_i) - \theta z_{ij}}{z_{ij}} \quad \forall i, j .$$

In the implementation, at first, preliminary values for Ψ_i are calculated

in the same way as for the MCMC. Secondly, the MCMC described in the previous section is run 100 times, with the only purpose of evaluating θ and z_{ik} . As far as the former, its last value is considered to be the best approximation, whereas for the latter, their values are averaged along the iterations. They are then used to construct the $\epsilon(t_{ij})$, as described in (3.3). The third and final step is the actual quantile regression. Because of the observations made about the quantile regression in a previous section, the model is run with $\tau = 0.65$ and tried for both single and grouped pixels. The resulting fitting can be seen in Figure 3.7. The two methods do not differ widely, but it is known that considering groups of pixels builds a more robust estimator.

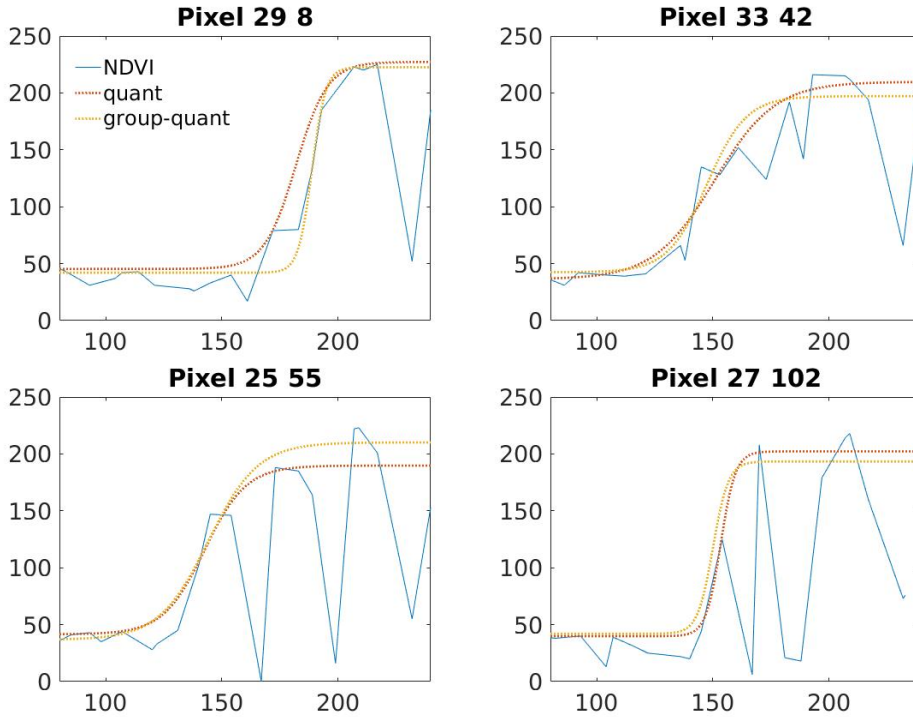


Figure 3.7: Fitting functions $f(t_{ik}, \Psi_i)$, obtained from quantile regression applied to single and grouped pixels, $\tau = 0.65$. The legend for the first pixel is valid for all four.

This last method appears to give the best fitting results. Furthermore, it is decided to rely on the estimation where groups of pixels are taken into account. The final improvement to test is therefore to use the function $g(\vec{t}_i, \Theta_i)$ on grouped quantile regression. The results for quantile regression on this new function are in Figure 3.8.

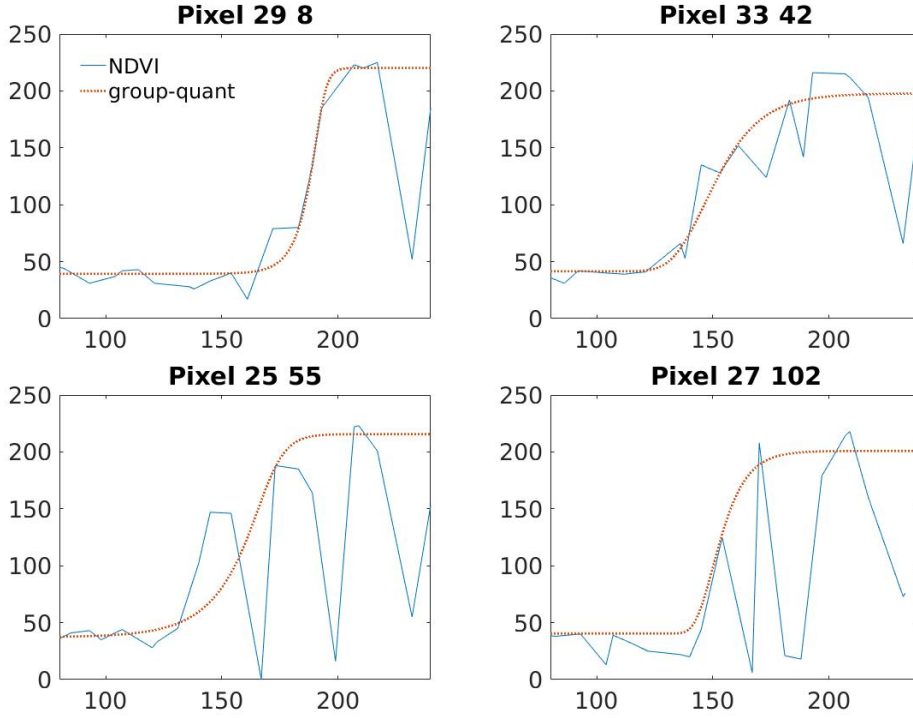


Figure 3.8: Fitting functions $g(t_{ik}, \Theta_i)$, obtained from quantile regression applied to grouped pixels, $\tau = 0.65$. The legend for the first pixel is valid for all four.

The relevance to include a further parameter in the estimation is checked by studying plots and calculating AIC and BIC for both $f(\vec{t}_i, \Psi_i)$ and $g(\vec{t}_i, \Theta_i)$. However, since the estimation of ν_i causes computational instability, and thus affects the estimation of AIC and BIC, the analysis is done keeping ν fixed for all pixels. The values for ν and the corresponding AIC estimations (BIC values were very similar) are summarized in the following table, where, as known, the second row corresponds to the function $f(\vec{t}_i, \Psi_i)$:

ν	AIC
0.1	$5.5 * 10^4$
1	$5.7 * 10^4$
2	$5.1 * 10^4$
2.5	$4.8 * 10^4$
3	$5.0 * 10^4$

According to these values, the best option appears to be $g((\vec{t}_i), a_i, b_i, t_{0i}, k_i, \nu_i = 2.5) \quad \forall i$.

Results and Observations

First, as it has been anticipated throughout the section, the quantile regression applied on a GH error structure is the best performing fitting model. In order to underline the relevance of the introduction of a more complex error structure, the values of z_{ij} , averaged along 100 MALA iterations are here plotted against the quality indicator of the measurements for the same pixel and time point. The resulting plot in Figure 3.9 shows how much higher z_{ij} are when the quality of the observation is null. This suggests that, even if the quality information was not taken directly into account, the new error structure is able to capture its variation.

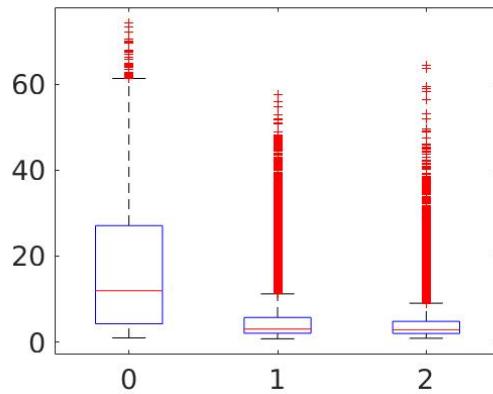


Figure 3.9: Distribution of z_{ik} depending on the quality of the observation.

Chapter 4

Near real-time detector

In order to recap, the best fitting found throughout the analysis that will be used in building the detector is:

- quantile regression applied to the GH error structure
- $\tau = 0.65$
- grouped pixels
- function $g(\vec{t}_i, a_i, b_i, t_{0i}, k_i, \nu_i)$
- $\nu_i = 2.5 \quad \forall i$

First of all, it is important to point out that all results plotted and discussed so far were calculated on the observations collected during the year 2000, the first year of the dataset used for this project. During that year no significant moth outbreak was registered. The years, among those in the dataset, when the forest around the Abisko national park was largely damaged by *Epirrita Autumnata* are 2004 and 2013. Therefore, the strategy is to build the defoliation detector based on the years 2000-2003, check it on 2004 measurements and then test it on 2013, since for this year the information about moth outbreaks is available. The idea for the detector is to build a robust model able to detect the occurrence of abnormal behaviour as early during spring season as possible. The parameter that will be the focus of the detector is a_i . Since it captures the NDVI level that is reached, it is able to detect the occurrence of abnormal behaviours.

4.1 Procedure

The detector is built on the fitting functions calculated between 2000 and 2003. The first step is therefore to estimate all parameters $(a_i, b_i, t_{0i}, k_i)_i$, for each year, with a quantile regression on a GH error structure, using grouped pixels.

Once the $(\Psi_i)_i$ are estimated, the properties of b_i , t_{0i} and k_i are studied. First, in Figure 4.1 one can see their distributions. Second, the correlation between parameters k_i , b_i and t_{0i} and forest fraction or altitude respectively is studied. Their possible correlation with altitude and forest fraction was intuitively seen in the plots of the quantile regression parameters in Figure 3.5. All combinations are checked, but only those in Figure 4.2 are significant. However, even among these, it appears that the relations between k and forest fraction and b and altitude are not highly significant, compared to the third where t_0 seems strongly positively correlated with the change in altitude. Intuitively, one can think that spring starts later on mountains' sides compared to valleys. This correlation will be used in the construction of the detector.

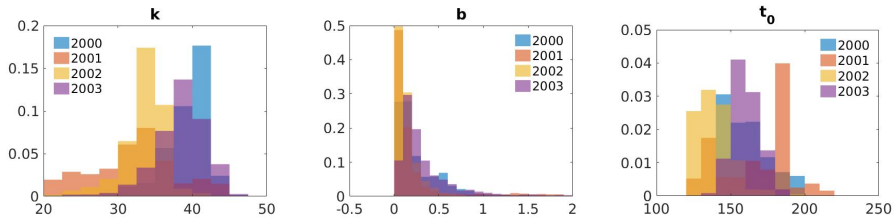


Figure 4.1: Distribution of parameters.

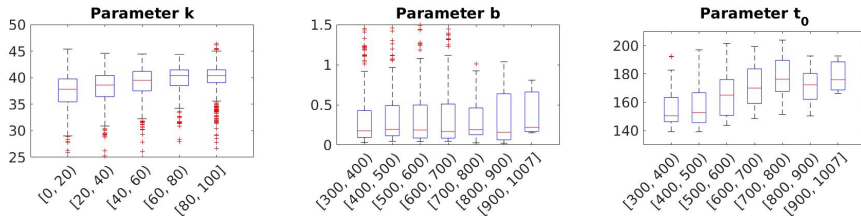


Figure 4.2: Correlation between k and forest fraction on the left and t_0 and altitude on the right.

The third step is different for each detector. The aim is to use information from previous years to improve the speed and accuracy of the fitting for 2004 data.

Once the model is defined, it is fitted to 2004. Since the detector is supposed to identify abnormal behaviours as soon as possible, all steps of the quantile regression are run for increasingly long intervals. At first, the observations are only four, in the interval 10-13 (from mid March to mid April); then two observations are added at each step until the last interval is the usual 10-30.

The detector is finally built according to the following idea. While the interval of provided observations grows, the parameter a_i is monitored. In particular, when most pixels have reached their high summer NDVI levels,

the attention focuses on those that have not. The details of this passage will be clarified for each detector.

4.1.1 Detector 1

In this first attempt, the amount of parameters that need to be estimated is reduced to one: a_i . This is achieved by using the properties of the other parameters that have been discussed previously. Since the goal is a near-real time detector, in the early spring, the model will rely on very few measurements in general and even fewer presenting quality higher than zero. Lowering the amount of coefficients that need to be found is then fundamental in order to get a good estimate. Two parameters are set equal to the median of their distributions for all pixels: b and k . As far as t_0 , its values for all years are regressed against altitude then, for each pixel i , the prediction is given to the model as a fixed value:

$$\alpha + \beta \times \text{altitude}_i = 130.34 + 0.06 \times \text{altitude}_i.$$

On the left side of Figure 4.3, a sample of the distributions of a_i , with larger and larger intervals, is presented. It is interesting to see how all values of a_i are very low when the intervals are short, and how they grow once the intervals get longer. This phenomenon can be seen on the fitting functions in Figure 4.4 as well. The final distribution that $(a_i)_i$ reaches, when the full interval $[10, 30]$ is provided, is very similar to the ones calculated for the years 2000-2003, as it can be seen on the right side of Figure 4.3.

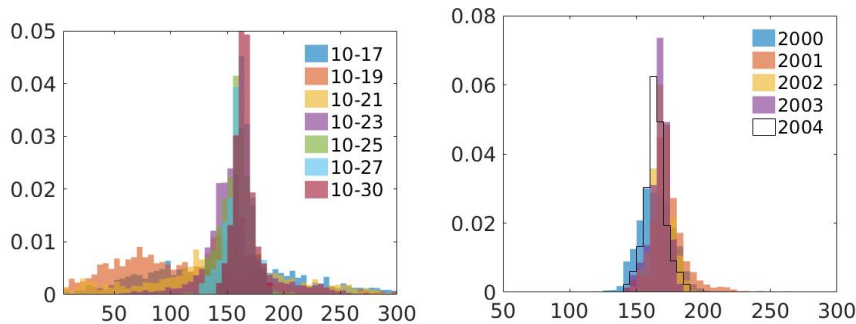


Figure 4.3: On the left, distributions of a_i with increasing interval of observations. On the right the final distribution of a_i for 2004 compared to the previous ones.

This similarity between distributions makes sure that distributions from previous years can be used to define an "acceptable" interval of a_i values, outside of which a moth outbreak could be happening. In Figure 4.4, pixel 29-8, seems to behave differently with respect to the other three, in particular, it has a slower growth. In fact, when looking at the estimation obtained

for interval 10-23, all other pixels reach their final a_i value, whereas 29-8 reaches it only when the provided interval is $[10 - 25]$. There is the possibility that this behaviour for the pixel 29 – 8 depends mostly on altitude, but it gives an idea on how to identify behaviours that lie outside the norm.

In conclusion, in order to identify abnormal behaviour, all a_i parameters evaluated for intervals $[10 - 21]$, $[10 - 23]$ and $[10 - 25]$ are tested against the common distribution for the years 2000-2003: if the a_i falls outside its 99% one-sided confidence interval then the i -th pixel is considered possibly affected by the moth.

However, among these pixels, those that are high on mountain sides show a late t_{0i} in general which could be mistaken for a moth outbreak. Therefore, for each interval $[10 - 21]$, $[10 - 23]$ and $[10 - 25]$, the maximum t_{0i} for which the method makes sense is roughly equal to $(\text{upper limit}) \times 8 - 10$. This relation is translated into an upper limit on the altitude, which is different for each observations interval that is provided to the estimation:

$$t_{0i} < \max(t_{0i}) \leftrightarrow \text{altitude}_i < (\max(t_{0i}) - 130.34)/0.06.$$

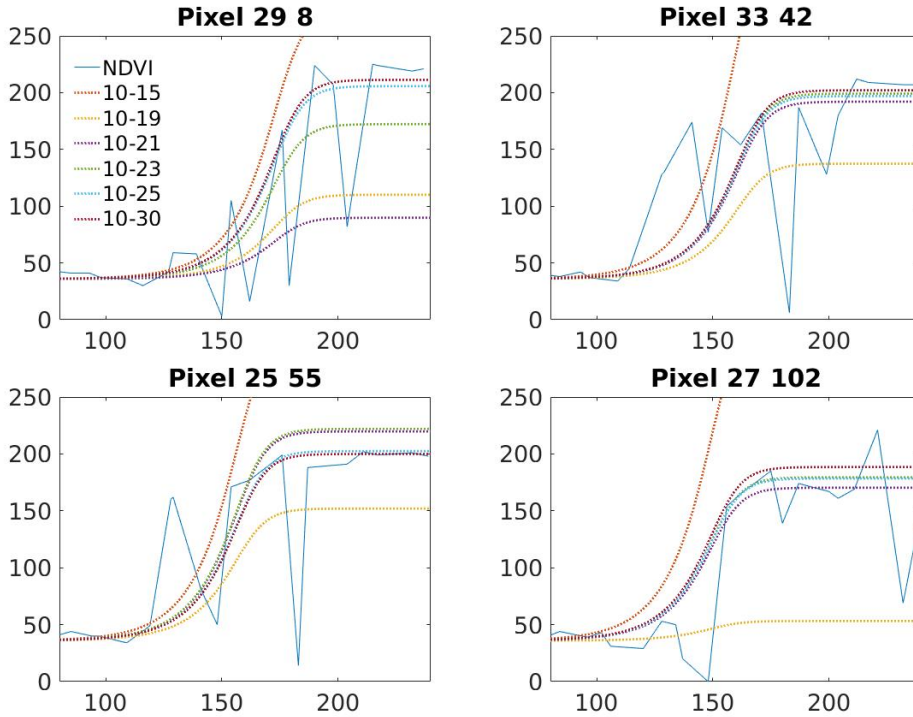


Figure 4.4: Fitting functions obtained from the detector. The interval of observations provided increases at each step, the legend of the first plot indicates the extremes of the interval at each step.

Results and Observations

In Figure 4.5 the pixels whose a_i might signal a moth outbreak are plotted in yellow.

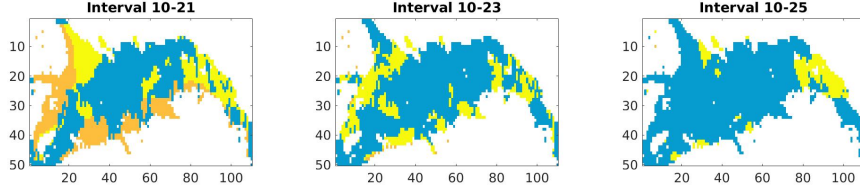


Figure 4.5: Possible locations of the moth in yellow, in orange areas that could suffer from moth outbreak or their delay could simply depend on altitude, in blue the rest. Results based on the first detector.

4.1.2 Detector 2

In the second attempt to build a detector, the parameters fitted for previous years are used to construct priors. The only fixed value is still $\nu = 2.5$.

In particular the function to minimize is the following:

$$\hat{\Psi}_i = \arg \min_{\Psi_i} \sum_i^{25} \sum_{j=1}^n \rho_{\tau}(\epsilon(t_{ij})) + \frac{1}{2} \left(\frac{(b_i - b_M)^2}{\sigma_b^2} + \frac{(k_i - k_M)^2}{\sigma_k^2} \right) + [\alpha \quad \beta] \Sigma [\alpha \quad \beta]^T.$$

The coefficients b_M , k_M , σ_b^2 and σ_k^2 are respectively medians and variances calculated from the values estimated for 2000-2003, the choice of the median depends on its robustness when compared to mean in the case of outliers. The coefficients α and β , instead, are the parameters obtained from regressing t_0 against altitude, the same presented for the previous detector.

The behaviour of the a_i parameter for this detector is highly similar to what was observed in the previous case. In fact, on the left side of Figure 4.6 it can be seen how the centre of the distribution grows together with the length of the interval. On the right side instead, its final distribution is comparable to the ones obtained in previous years. This allows for a detection with a procedure similar to Detector 1. It is however interesting to notice that this Detector appears to be slower. Having 5 parameters to evaluate for each pixel, at each step, the model needs more observation in order to reach a good estimation of a_i . This can be read when comparing the left side of Figure 4.3 and 4.6, it seems therefore better to use the intervals $[10 - 23]$, $[10 - 25]$ and $[10 - 27]$ in this case. The altitude minimum is set in the same way as before.

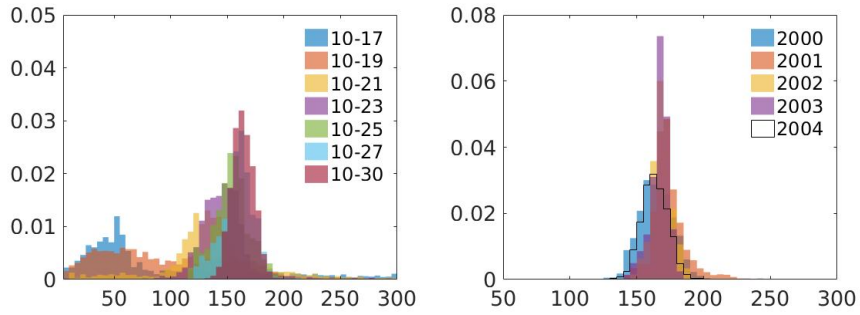


Figure 4.6: On the left, distributions of a_i with increasing interval of observations. On the right the final distribution of a_i for 2004 compared to the previous ones.

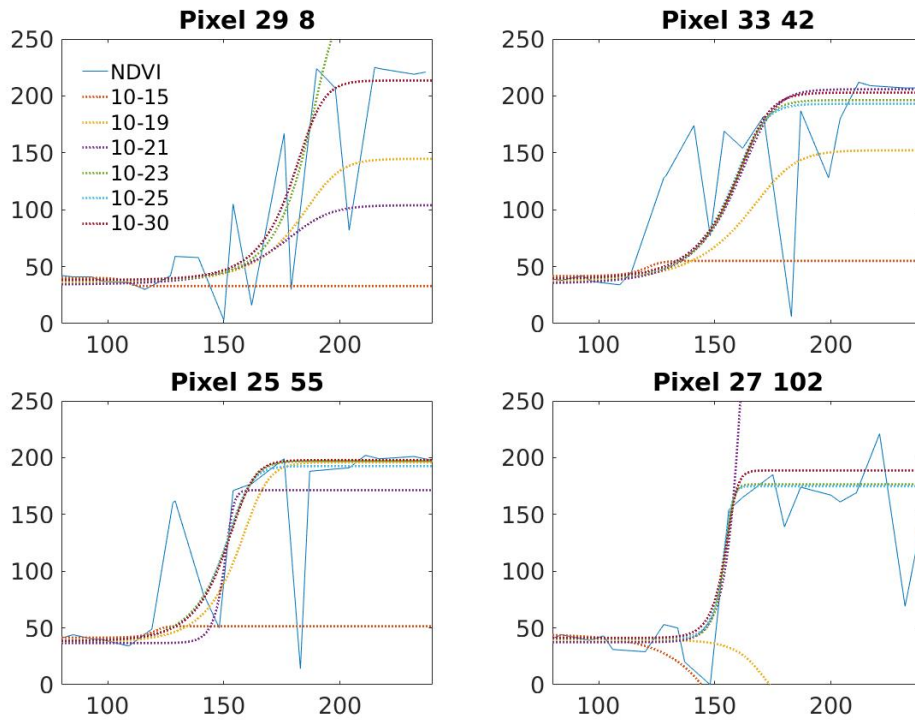


Figure 4.7: Fitting functions obtained from the second detector. The interval of observations provided increases at each step, the legend of the first plot indicates the extremes of the interval at each step.

Results and Observations

In Figure 4.8 the pixels whose a_i might signal a moth outbreak are plotted in yellow.

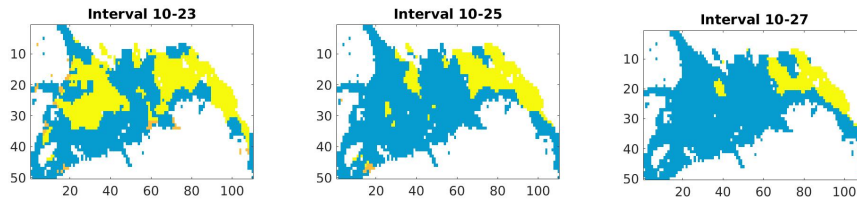


Figure 4.8: Possible locations of the moth in yellow, in orange areas that could suffer from moth outbreak or their delay could simply depend on altitude, in blue the rest. Based on the second detector.

4.1.3 Detector 3

In this third attempt a combination of the previous two is tried out. Parameters b and k which are less influential and do not vary widely throughout the park, are kept fixed for all pixels, their value equal to the median of their respective distributions for the years 2000-2003. The parameter t_0 is instead evaluated for each pixel through the regression on altitude. Therefore, the function to minimize is the following:

$$\hat{\Psi}_i = \arg \min_{\Psi_i} \sum_i^{25} \sum_{j=1}^n \rho_{\tau}(\epsilon(t_{ij})) + \frac{1}{2}([\alpha \ \beta] \Sigma [\alpha \ \beta]^T).$$

This third detector appears to be similar to the first one, as far as the speed at which a_i grows, the intervals used for the detection are therefore $[10 - 21]$, $[10 - 23]$ and $[10 - 25]$.

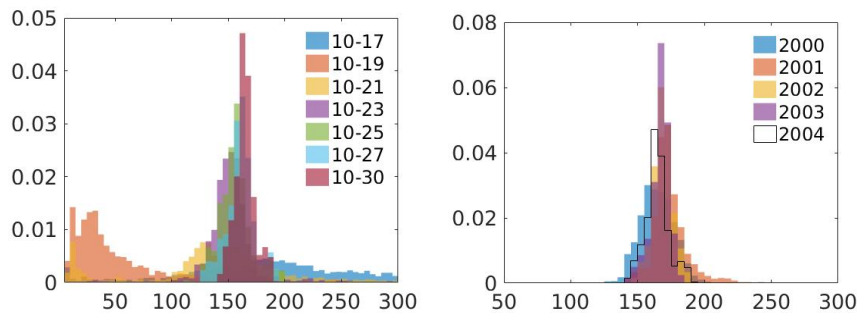


Figure 4.9: On the left, distributions of a_i with increasing interval of observations. On the right the final distribution of a_i for 2004 compared to the previous ones.

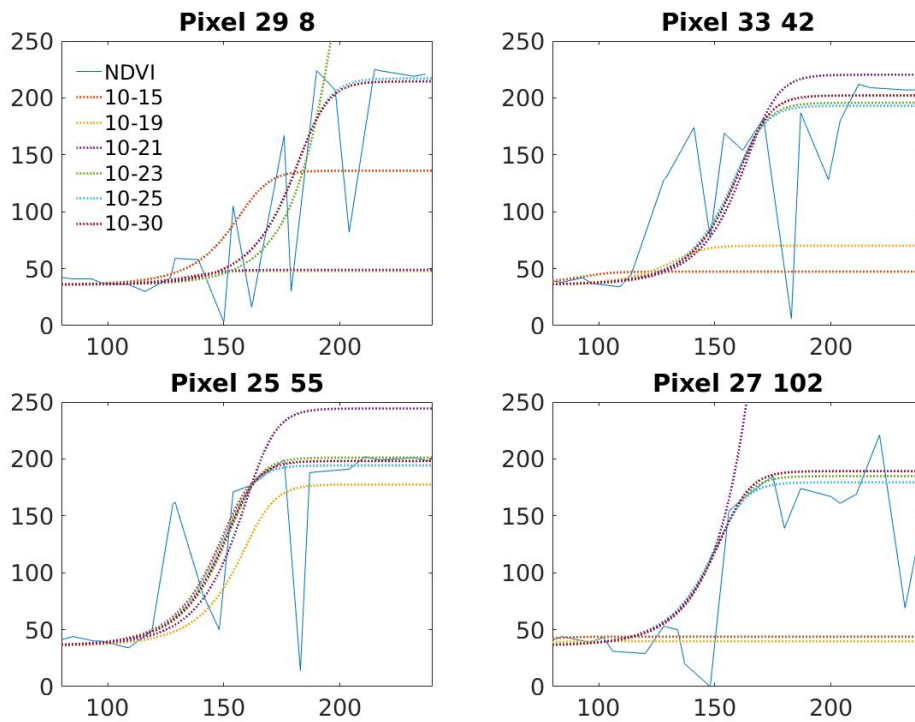


Figure 4.10: Fitting functions obtained from the second detector. The interval of observations provided increases at each step, the legend of the first plot indicates the extremes of the interval at each step.

Results and Observations

In Figure 4.11 the pixels whose a_i might signal a moth outbreak are plotted in yellow.

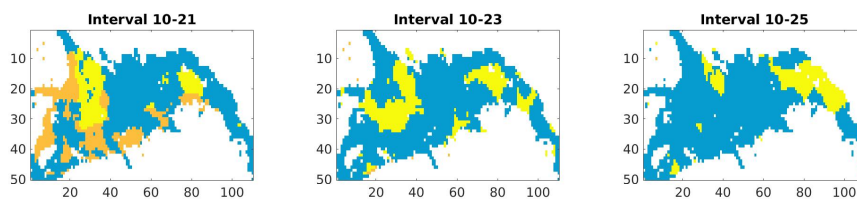


Figure 4.11: Possible locations of the moth in yellow, in orange areas that could suffer from moth outbreak or their delay could simply depend on altitude, in blue the rest. Based on the third detector.

4.1.4 Comparing detectors

Comparing the figures of the detected zones, it appears that all detectors agree on the North-eastern area on the shores of the lake and a small portion

in a central-North position. The second detector reaches this conclusion only on the interval $[10 - 27]$, while the other two act faster, it is therefore discarded for the next step.

Between the first and the last detector, the latter seems preferable since t_0 is estimated through the regression coefficients. It is known the parameter is affected by temperature and could globally vary from one year to another. The computational trade-off is low, and the possible gain in accuracy is worthy.

4.2 Check detector on 2013 data

The assumption for this last part is that the observations for 2013 are given day by day, in order to really test the ability of the detector. The estimation is thus run for different intervals of observations following the procedure of the third detector.

Further assumption is that the distribution of the a_i parameter at the end of 2013 is similar to those between 2000 and 2003 and that, while the interval grows in length so does the average value for a_i .

Finally, the intervals chosen for the detection are $[10 - 23]$ and $[10 - 25]$, in order to see how the possible affected areas change on the map. In Figure 4.12, the areas where, according to the detector, a moth outbreak could be occurring.

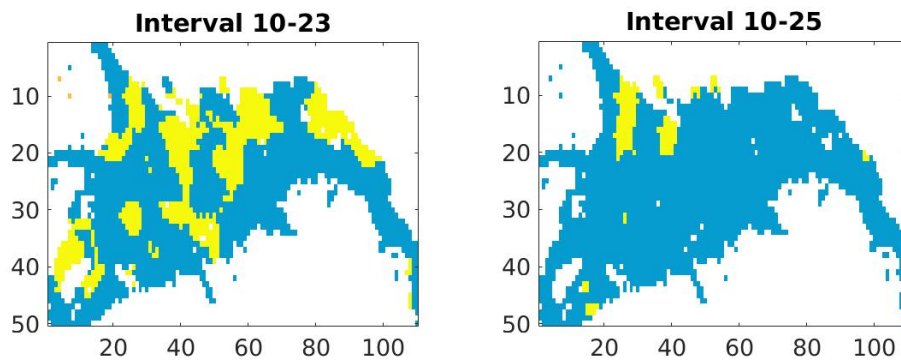


Figure 4.12: Possible locations of the moth in yellow, in orange areas that could suffer from moth outbreak or their delay could simply depend on altitude, in blue the rest. Estimation for 2013, based on the third detector.

4.2.1 Actual outbreak

In the final Figure 4.2.1, actual field information from a few field sites regarding a 2013 moth outbreak, is plotted. Those pixels where the moth outbreak is known to have happened are depicted in yellow, while in orange

one can see the pixels for which it is known that no damage occurred. No information is available for the other pixels throughout the park.

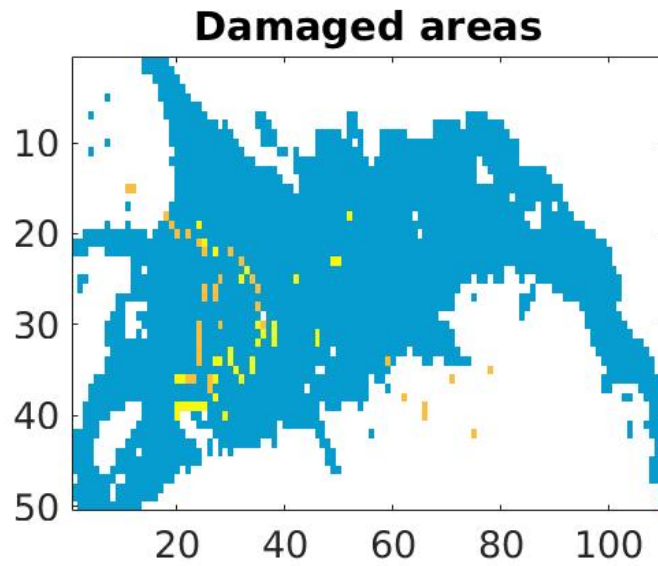


Figure 4.13: In yellow pixels known to be affected by the moth outbreak of 2013, in orange those that are known not to be affected.

Chapter 5

Conclusions

5.1 Models

The main observation about the fitting models is that the first simplistic approaches, minimizations done with LS and LAD, do not perform well in an analysis such as this one, where the noise affecting the measurements is high and frequent. The low quality of the data, readable in Figure 2.2, suggested from the beginning that strong measures against noise needed to be taken into account. It can therefore be concluded that those first approaches were obviously ineffective and could have been avoided, allowing more time to the more sophisticated tools.

5.2 2013 Detector

In the final chapter, after the detector is applied to the data from 2013, it appears that the areas in Figure 4.12 are much wider and do not correspond to the ones where the outbreak was registered by in-situ observation.

		actual		
		yes	no	no-info
predicted	yes	10	10	706
	no	27	18	1770

This discrepancy would suggest the detector is ineffective, it is therefore further investigated by plotting the NDVI of four pixels in Figure 5.1. They are, in order, a true positive, a false positive, a false negative and a true negative. They are plotted against the changing fitting function, in order to fully understand how the detector interpreted their behaviour.

As far as the true positive 19 – 24, its behaviour is exactly what would be expected from a pixel suffering from moth defoliation. The false positive 18 – 18 instead, could be the combined result of two factors: high noise between the days $\sim 170 - 190$ and altitude of 690m that forced t_0 to be too

high because of the regression. In fact, it can be seen how the first high values are ignored, the noise affects the observations at the end of June and the fitting function is only able to capture the summer values, reporting it as a delayed bud burst. As far as the false negative 18 – 52, the signal does not appear corrupted at all, the NDVI behaves very similarly to the true negative 21 – 24.

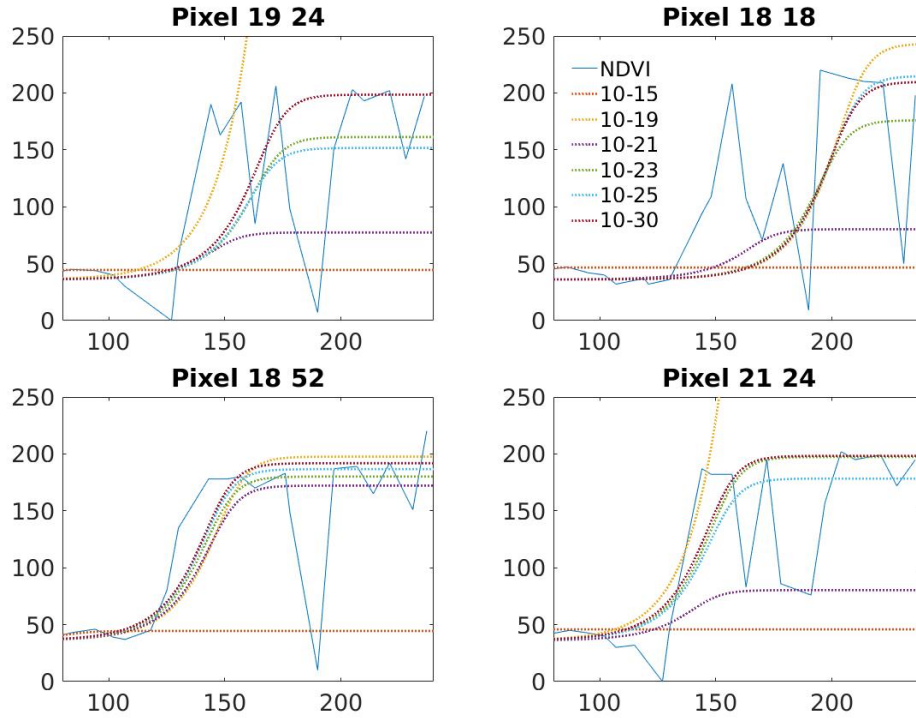


Figure 5.1: NDVI values for four pixels affected by the moth outbreak of 2013.

The behaviours of these four pixels and the others in their respective categories, suggested the following improvements for a follow-up research.

- The fixed dependence of t_0 on altitude may be too rigid, as it appears to happen for the false positive. Counter-measures were taken in order to avoid false positives whose delay depend solely on altitude, but they could be improved. Other solutions could be to include more covariates in the regression, daily temperature could be a good candidate, or simply allow t_0 to be estimated on its own.
- Having both b and ν constant could be too strict. The shape of the fitting function is identical for all pixels, it is only allowed to shift depending on t_0 and a . Therefore allowing b to be estimated at each time-point and pixel could be worth a try.

- NDVI may be capturing tree species other than birch, within the same pixel. Other trees may not be affected by *Epirrita Autumnata*'s larvae, thus making it impossible for the detector to distinguish between true and false negatives throughout the park.

Bibliography

- [1] SCB, “Exports by important site commodity groups,” 2017. Accessed: 2017-06-07.
- [2] C. Rullan-Silva, A. Olthoff, J. Delgado de la Mata, and J. Pajares-Alonso, “Remote monitoring of forest insect defoliation -a review-,” *Forest Systems*, no. 3, pp. 377–391, 2013.
- [3] P.-O. Olsson, J. Lindström, and L. Eklundh, “Near real-time monitoring of insect induced defoliation in subalpine birch forests with modis derived ndvi,” *Remote Sensing of Environment*, vol. 181, pp. 42 – 53, 2016.
- [4] L. Torbjörn, 2017.
- [5] “Abisko national park.” <http://www.nationalparksofsweden.se/choose-park---list/abisko-national-park/>, 2017. Accessed: 2017-02-20.
- [6] “Abisko scientific research station.” <http://polar.se/en/abisko-naturvetenskapliga-station/>. Accessed: 2017-02-20.
- [7] H. Bylund, “Climate and the population dynamics of two insect outbreak species in the north,” *Ecological Bulletins*, no. 47, p. 54, 1999.
- [8] U. S. Geological Survey, “Landsat earth observation satellites,” 2016.
- [9] T. Garcia-Mora, J. Mas, and E. Hinkley, “Land cover mapping applications with modis: a literature review,” *International Journal of Digital Earth*, vol. 5, no. 1, pp. 63 – 87, 2012.
- [10] E. S. Agency, “Terra/aqua modis,” 2017.
- [11] L. Eklundh and P. Jönsson, *Remote Sensing Time Series*, ch. 7. Springer International Publishing Switzerland, 2015.
- [12] A. Bannari, D. Morin, F. Bonn, and A. R. Huete, “A review of vegetation indices,” *Remote Sensing Reviews*, vol. 13, no. 1-2, pp. 95–120, 1995.

- [13] J. J. W. Rouse, “Monitoring the vernal advancement and retrogradation (green wave effect) of natural vegetation,” 1973.
- [14] S. Rawlings, J.O. Pantula and D. D.A., *Applied Regression Analysis*. Springer, second ed., 2001. pp.507-508.
- [15] R. Koenker, *Quantile Regression*. Cambridge University Press, 2005. pp.7,42.
- [16] D. Kraft, “Reconstruction of ndvi data with normal variance-mean mixture noise using stochastic gradient methods,” 2017. Student Paper.
- [17] O. E. Barndorff-Nielsen, “Normal inverse gaussian distributions and stochastic volatility modelling,” *Scandinavian Journal of Statistics*, vol. 24, no. 1, pp. 1–13, 1997.
- [18] W. R. Gilks and P. Wild, “Adaptive rejection sampling for gibbs sampling,” *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, p. 337, 1992.
- [19] G. O. Roberts and O. Stramer, “Langevin diffusions and metropolis-hastings algorithms,” *Methodology & Computing in Applied Probability*, vol. 4, no. 4, pp. 337 – 357, 2002.
- [20] M. Girolami and B. Calderhead, “Riemann manifold langevin and hamiltonian monte carlo methods,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 73, no. 2, pp. 123–214, 2011.

Abstract

Early and reliable detection of pest infestation in forests is crucial to protect the health of trees. Insect outbreaks are an important cause of defoliation: they delay the blossoming of leaves and thereby affect the growth of trees. Such delays lower the economic potential of forests as well as their capacity to absorb atmospheric CO₂. The aim of this thesis is to use mathematical models to build a near real-time detector that can identify abnormal behavior in forest growth, which can be the sign of pest infestation. In order for the detector to provide effective early warnings, the analysis uses high time-resolution satellite images.

In particular, the detector is based on data collected at the Abisko National Park in the north of Sweden. The insect that is known to threaten this birch forest is a moth called *Epirrita Autumnata*, whose larvae feed on the bursting leaves early in the spring.

The data used to study the behaviour of green mass is the Normalized Difference Vegetation Index (NDVI). A dataset of 14 years of measurements was available over an area of 350 km². The index is calculated based on images collected by the MODerate resolution Imaging Spectroradiometer (MODIS), which is placed on two satellites that orbit Earth and collects images of the surface daily. The analysis also includes information on forest fraction and altitude for each pixel of the area.

The first step of the analysis is fitting a function to the NDVI measurements of each pixel. The chosen function captures all important aspects of the change in NDVI during the spring. Different methods are used to fit the function to the NDVI. The first, simplistic models fail to fit the function because of the strong noise that affects the measurements. Therefore more robust and complex estimators are tried out. The final, best performing technique is used to construct the detector.

The idea behind the detector is to identify those pixels for which NDVI grows slower than expected, based on the values of other pixels and previous years, during the spring. In particular, it is known that, within the available dataset, the two years that have suffered a moth outbreak are 2004 and 2013. Hence, the chosen function is fitted to the data from 2000-2003, in order to get a sense of the behaviour of NDVI during "healthy" years. The detection of abnormal behaviour is done for 2004 and it is then tested on 2013. For this last year, a few locations of the outbreak were known, so that the results generated by the detector could be verified. The discrepancy between field data and the results generated by the detector suggests further adjustments that would improve the capacity to detect moth infestation.

Master's Theses in Mathematical Sciences 2017:E43
ISSN 1404-6342
LUNFMS-3070-2017
Mathematical Statistics
Centre for Mathematical Sciences
Lund University
Box 118, SE-221 00 Lund, Sweden
<http://www.maths.lth.se/>