A quantile regression analysis on the impact of climate change on the seasonal pollen release in Sweden

Max Brehmer

2023-05-11

Abstract

Abstract kommer...

Preface

Här tänkte jag skriva tack till handledare, en acknowledgement om att jag använt AI verktyg som ChatGPT och övrig extra information om uppsatsen.

Introduction

Over the past couple of decades it has been made clear that our climate is changing rapidly in various ways. Most notably the global average temperature has risen by ca $0.2°C$ per decade since the mid seventies, this constitutes an almost $1°C$ increase over the past half century [@hansen2006global]. Limiting our view to only Sweden, we also see significant shifts in this regions' otherwise stable climate. Several studies in recent decades draw the conclusion that plant phenology is impacted by this increase in temperature. [@van2002influence] discusses in their study of the seasonal pollen shift in the Netherlands that an advance in the start of the pollen season by 3-22 days took place in the latter third of the 20th century. Likewise this paper strives to understand what seasonal changes have occurred to the pollen season in Sweden. As mentioned in [@lind2016pollen] the results may differ for various species of pollen. More precisely they found a stark difference in duration among arboreal plant species compared to herbaceous ones, with the former trending towards an earlier end date, while the latter was pushed to a further date and thus have a longer seasonal duration. Grass pollen, being herbaceous, is the leading cause of pollen allergy in many developed countries, meaning a lot of people suffer from these seasonal changes for an extended time [@garcia2017poaceae]. In Sweden and other parts of northern Europe however, due to differences in temperature and overall climate, the arboreal types like birch (betula) are the most common cause of pollinosis [@d2017allergenic].

Continous monitoring of pollen conducted by the Swedish Museum of Natural History (NRM) began in 1973 at the Palynological laboratory in Stockholm. Since then multiple other stations have been included in the scope of NRM's continous pollen monitoring program. As of 2022 there are 20 active stations involved [@nrm2022pollen], monitoring the release of several unique species of pollen. In this paper we consider 7 of the most allergenic species, these are the arboreal pollen of alder (alnus), birch (betula), hazel (corylus), oak (quercus), willow (salix), elm (ulmus) and the herbaceous species of grass pollen (poaceae).

In this paper we will attempt to determine the shift in dates of the start and end of the pollen season in Sweden as an effect of global warming of Earth's climate. We will consider global warming as a linear trend over the researched time period as to simplify the process of analyzing pollen patterns. We can do this as research has shown acceptable fitting of linear models over anthropogenic climate change, in regard to temperature [@hansen2006global]. An analysis will be conducted based on two separate frequentist quantile regression models, namely linear regression on empirical quantiles (EQ) and non-parametric quantile regression (QR). In a study of seasonal shifts of migratory birds [@karlsson2022comparison] perform both these methods, this

paper covers the majority of the theory in regards to the construction of the statistical models. In the case of QR, we also make good use of [@takeuchi2006nonparametric] for a more in depth description of the method.

By its conclusion this research paper aims to have built a statistical model that can explain the historic shift in pollen seasons for each of the 7 species and also possesses the ability to predict expected further changes in the Swedish pollen season.

This paper begins with the **Litterature review** section by presenting the findings of what we consider to be the most relevant studies of similar character to what we envisage our own research to look like.

The **Data** section is where all aspects of the data used in the research is presented. We cover these aspects in three smaller sections. Firstly **Data collection** where a light description of both the physical and digital methods used to collect the data. An extensive description of how we use the data is later presented in the section **Understanding the data**. Here the reader is granted a look at the structure of the data in the form of easily-to-read tables and which variables are to be considered in the research. We conclude the **Data** section with **Data selection** where an explanation of how we deal with missing values is conducted.

Section 3, **Methods** encompasses all the theoretical definitions of the methods included in this paper. In order to strengthen the significance of our results, we have used two independent methods. Linear regression on empirical quantiles and Nonparametric quantile regression. Linear regression on empirical quantiles, which we refer to as EQ, is a simple method commonly used to perform statistical analyses on different quantile levels of data, while nonparametric quantile regression (QR) is a more complex method which substitutes the conditional mean of the loss function for the conditional median and requires numerical approaches to the resulting optimization problem. Our sections on EQ and QR respectively explain the theoretical aspects of these methods, while simultaneously providing descriptions of how we have applied them in our models.

A section containing the most important results of our models is given in **Results**. This part of the paper is divided in to 2 sections for increased readability. **Plots** ??? contains visual descriptions of how the pollen season behaves over time, for selected combinations of location and species. The remaining figures are available in the appendix. **Tables** is where the key values are presented in the form of ANOVA tables and alike. These values include regression coefficients, p-values and R^2 for each model. Again, as for the plots, results for all combinations of data are not presented in this section. Rather a selection of the most significant, representative and most extreme results are, while remaining tables can be found in an appendix.

The **Discussion** is the section in which we attempt to make sense of the results observed and visualized in the **Results** section. For each of the analyzed results we may be able obtain a definitive conclusion, or not, in which case a description of what we are missing for us to draw a conclusion follows. Since there are limits to what we can expect to answer in this thesis, we have set aside a section (**Further improvements**) on how, and which parts of our work can be improved upon.

This paper concludes with an **Appendix** section. Additional or complementary information which the reader may find interesting, but not vital for the results of which we form our conclusions. Each appendix is reffered to by, and refers back to a result or explanation.

Section 1: Litterature review

I mån av tid skriver jag lite om vad som tidigare uppmärksammats i liknande tidigare rapporter.

Section 2: Data

Section 2.1: Data collection

As we mentioned earlier in this paper, the monitoring of pollen in the Stockholm region is conducted by the Palynological laboratory at NRM. The laboratory in question uses a Burkard Seven Day Volumetric Spore Trap to capture pollen and spores from the air through a small entrance meant to resemble the human airways. Thus approximately 10 liters of air passes through the machine each minute, which is what humans tend to consume. In order to capture the pollen particles carried by the passing air, a sticky tape is mounted to a drum rotating at 2 mm per hour. As only a small portion of the tape is exposed

to the air at each point in time, this method grants us a good indication of the volume of pollen in the passing air at any given moment. Each captured pollen is individually counted with regular intervals using microscopes. It must be noted that not all stations possess the same equipment. In particular, differing microscope sizes are used across the country. Consequently, the measured values of the pollen counts are biased towards the larger microscopes, thus showing a somewhat inaccurate representation of the true pollen counts [@nrm2022microscope] [@nrm2022pollen2]. However considering the structure of the dataset and the consequent data analysis being relativistic, for which a descriptive presentation follows in the **Data** section, this phenomenon has been ignored.

Section 2.2: Understanding the data

The dataset that we have at our disposal contains 5 unique variables: `date`, `station`, `name`, `count` and `factor`. Of which all but the `factor` variable are used in this research paper. We have also added a `latitude` variable since it is known that higher latitudes contribute to more extreme climate changes [@alecrim2023higher]. This variable is however entirely dependent on `station`. A short description of the meaning of each variable is shown in table 1. Equipment for data collection is active only during the period of the predicted pollen season, estimated through using predictive models for pollen activity based on historic results.

Table 1: Description of the variables present in the datasets used in this research paper.

Variable	Type	Decription
Station	categorical	Geographic location of the pollen monitoring station.
Species	categorical	Genus of the recorded pollen counts.
Date	continous	Gregorian calendar date on which the airborne pollen were registered.
Count	continous	Number of individual pollen counted.
Factor	continous	Reference variable for the size of the microscope used.
Latitude	continous	Northern latitudinal cooardinates of said station.

Viewing the years of availability in figure ? we conclude that not all data points are present in the data set. The stations had differing opening dates and not all species tend to be available to begin with. If no consideration for the location of said data points are made, we may observed skewed results, as the geographic distribution of monitored pollen changes over time due to availability. Thus analyzing the data in geographic categories of where they were collected is a necessary consideration.

Table 2: Years for which data for different pollen species are available at the pollen monitoring stations used in our research, as well as their latidudal coordinates.

Station	Latitude	Pollen genus (since ...)
Umeå	62.83	Alnus, Betula, Poaceae, Ulmus (1979), Salix (1981), Corylus (1987), Quercus (1995)
Eskilstuna	59.37	Alnus, Betula, Corylus, Poaceae, Quercus, Salix, Ulmus (1976)
Stockholm	59.33	Alnus, Betula, Corylus, Poaceae, Quercus, Ulmus (1973), Salix (1977)
Norrköping	58.59	Alnus, Betula, Corylus, Poaceae, Quercus, Salix, Ulmus (1987)
Jönköping	57.78	Alnus, Betula, Poaceae, Quercus, Salix, Ulmus (1988), Corylus (1989)
Västervik	57.76	Alnus, Betula, Corylus, Poaceae, Quercus, Salix, Ulmus (1987)
Göteborg	57.71	Alnus, Betula, Corylus, Poaceae, Quercus, Salix, Ulmus (1979)
Malmö	55.60	Alnus, Betula, Corylus, Poaceae, Quercus, Salix, Ulmus (1979)

Section 2.3: Data transformation

In order to be able to perform methods like quantile regression and linear regression on quantiles, a transformation of the given data structure is necessary. To be more precise, we must turn each individual

pollen, quantified by the `count` variable, in to its own data point. Using the function `uncount()` in `R` we are able to perform the described transformation of data. The daily counts of certain pollen types at any given station can be in the hundreds. Each pollen season usually lasting around a month, and the data covering multiple species of pollen at a multitude of geographic locations over a number of years, means in return, that the size of the dataset increases dramatically and as a consequence slows down computation time.

Section 2.4: Data selection

After altering the structure of the data in to an individualistic format, we first remove any observations with missing values in any of the columns `date`, `station` or `species`, since these are fundamental parameters to perform the following data analysis. For the purposes of EQ regression, the time it takes to compute the models on the entire dataset in `R` is negligible. For nonparametric quantile regression however, the computational time appears to grow very quickly with the amount of observations. This has led us to the decision to reduce the amount of content in each dataset (one dataset containing all observations of any combination of `station` and `species`) to below a fixed limit. Without reducing any data, the largest data sets contain ~ 500000 observations, while some combinations of data contain considerably fewer. By testing and optimizing the models for various sizes of the datasets we decided to set the limit to 5000 observations per combination of data. This reduces the running time considerably, making the process of analyzing our results much easier. In order to minimize the affect this has on our results the rows must be removed uniformly over the time of year. If the number of observations in a dataset is smaller than $n \cdot 5000$, we want to select every nth element and discard the rest. We can do this in `R` by using the `slice()` and `bind_rows()` functions from the `dplyr` package. By performing these modifications to our data, we are sure to keep the shape of the distributions the same and thus the results of the quantile regression.

Section 2.5: Annual distribution of pollen

Before constructing the models, we can take a glance at what sort of results to expect from the subsequent regression analysis. In figure (?) we present a ridge plot containing information about the distribution of observed pollen by date of year. The `geom_density_ridges()` function from the `R` library `ggridges` uses kernel density estimation (KDE) to fit a continuous density function to what is, in essence, a histogram of the frequency of pollen on each date. KDE is a non-parametric method used to estimate the probability density function of a random variable. This is achieved by placing a kernel at each data point. In our case the kernels are Gaussian distributed. We then summarise each of the kernels' distributions in order to obtain a smooth continuous approximation of the density. The width of the kernel, known as the bandwidth, determines by how much each data point contributes to the density estimate. Larger bandwidths result in smoother but less detailed density estimates, while smaller bandwidths result in more detailed but noisier density estimates [@mvstat2001kde]. A common way to find the optimal bandwidth is to use the Asymptotic Mean Integrated Squared Error (AMISE). The function `geom_density_ridges()` uses an alternative method based on the Sheather and Jones (1991) plug-in bandwidth selection method. [@10.2307/2345597]

By looking at the average distribution from the first and last 5 years in the dataset respectively, we construct a somewhat consistent average in terms of the state of the pollen season at a given location, reducing the risk of falsely identifying outlier data as patterned occurrences. We can conclude that, in the Stockholm area, the pollen season seemed to begin earlier for all the arboreal pollen species while grass pollen (Poaceae) did not appear to show those tendencies. In general the shape of the pollen distribution was not seen to have been altered over time. A visual analysis of the distribution of annual pollen release may suggest that the average pollen season now begins earlier by up to a month compared to 45 years ago.

Section 3: Methods

In this section we explain the underlying theory behind the statistical approaches used for our data analysis. We will compare the simpler "empirical quantile linear" model (EQ), as used in [@karlsson2022comparison] with the supposedly more powerful method "nonparametric quantile regression" which we will refer to as QR.

By the QR approach, we estimate the response variable as the conditional median of the predictor variables, of which the median can be substituted to any other quantile of data. In the EQ method however, we use the

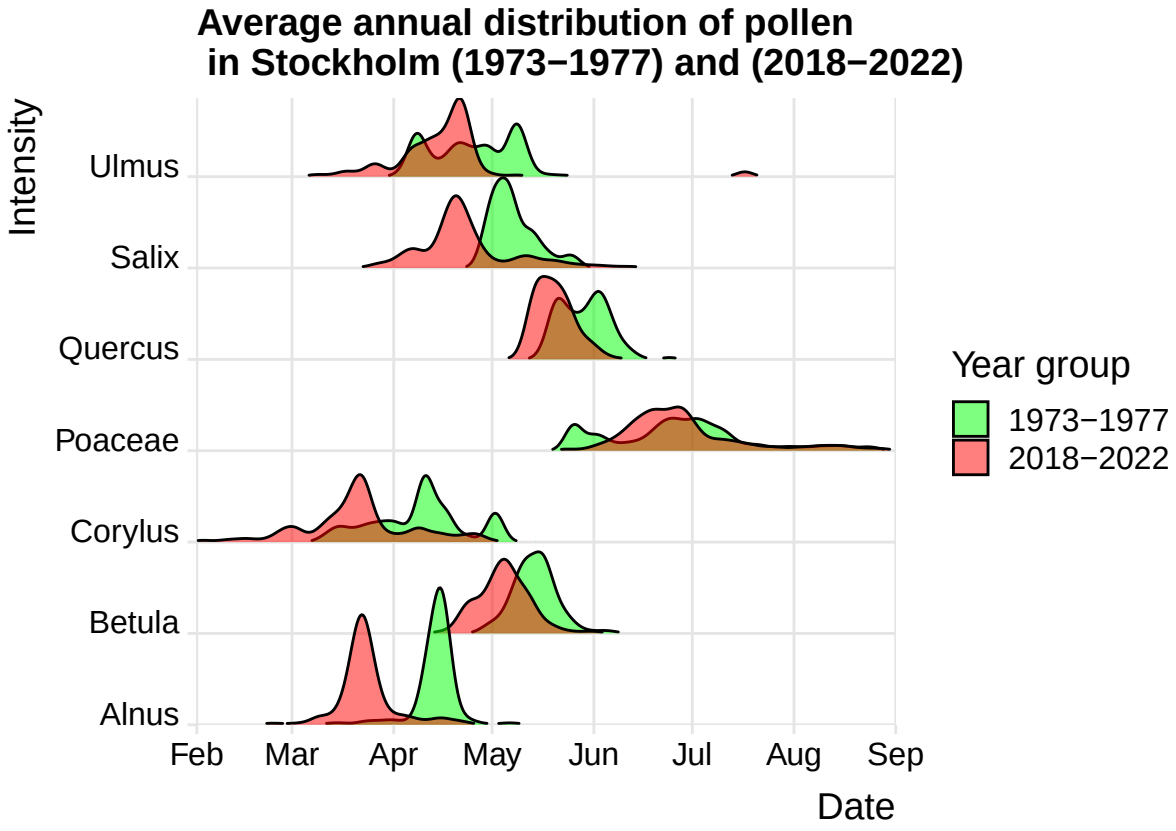

Figure 1: Ridgeline figures approximating the density functions of the distribution of annual pollen. Comparing the observed dates of pollen capture for the average of the years 1973-1978 (green) with the average for 2018-2022 (red).

method of ordinary least squares (OLS) to estimate the conditional mean of a subset of the predictor variables corresponding to the desired quantile level [Wikipedia]. Ordinary linear regression is the preferred method for many research purposes due to its inherent simplicity. In this research paper however, we are more interested in patterns for certain quantiles of data than the mean. To do this effectively we have conducted our research using quantile regression methods instead.

An aspect in which quantile regression performs better than linear regression is when data is not homoscedastic or normally distributed. Linear regression models perform poorly for data that is heteroscadastic and/or non-normally distributed. By estimating the conditional median however, as opposed to the mean, we get a model that yields better predictions for data with these properties. Another advantage of using quantile based methods is that they are less sensitive to outliers, since nonlinear tendencies may lead to abnormal behaviors for more extreme observations, which can be more accurately accounted for when looking at quantiles rather than the mean.

All models in this paper will use `date` as the response variable, or a yearly quantile of `date`. The release date of pollen grain $i \in \{1, ..., n\}$, where n is the total amount of pollen grains released, is predicted by the models as the response y_i. We form a covariate vector $x_i = (1,\, t_i)$ which includes an intercept and the predictor variable $t_i \in \mathcal{T} = \{1, ..., T\}$ which refers to the gregorian year of observation with T being the amount of years monitored.

In order to make an educated analysis of the behaviour of pollen release we need a strict definition for what constitutes a pollen season. There have been many attempts by various authors to find an optimal definition for the dates of which each pollen season covers. In this paper we classify the dates within which the pollen season is deemed active by referring to the EAN definition. [bastl2018defining] The EAN database contains a lot of data on pollen release in which they define the pollen season as the date at which at least 1% of the annual pollen have been counted. By the same definition, the pollen season is said to end when 95% of all pollen has been released. A reason as to why the starting quantile level and the one at the end of the season are not symmetric (i.e. why $Q_{start} \neq 1 - Q_{end}$) may be similar to how infection transmission models tend to behave, namely that the beginning of the season tends to be very intense while the distribution decreases more slowly towards the end of the season.

When performing calculations of the arithmetic mean of a vector, we use the following formula:

$$\bar{x} = \arg\min_{\mu \in \mathbb{R}} \sum_{i=1}^{n} (y_i - \mu)^2,$$

where y corresponds to the vector of observed values and μ is a scalar value that minimizes the sum of squares.

If we instead want to calculate the arithmetic median, we use the absolute values between the vector of observations and the average instead of the square of these values. The median is thus provided by

$$\hat{x} = \arg\min_{\mu \in \mathbb{R}} \sum_{i=1}^{n} |y_i - \hat{x}|.$$

Furthermore, if one wants to compute any given quantile τ given a vector of observations, one may use the minimization algorithm

$$x_\tau = \arg\min_{\mu_\tau \in \mathbb{R}} \sum_{i=1}^{n} \mathcal{L}_\tau(y_i - \mu_\tau),$$

where $\mathcal{L}_\tau(\xi)$ represents the pinball loss function, introduced by Roger Koenker in [@koenker2005qr] (Source mut be older?). This is a convex, piecewise linear function which computes the deviation between the predicted quantile and the actual value of the target variable. It is defined by

$$\mathcal{L}_\tau(\xi) = \begin{cases} \xi \cdot \tau & \text{if } (\xi) \geq 0 \\ \xi \cdot (\tau - 1) & \text{if } (\xi) < 0. \end{cases} \qquad (?)$$

The name of the pinball loss function derives from the fact that it's shape somewhat resembles that of a pinball game as shown in figure (?). Rather trivially, the loss is minimized at $\xi = 0$, since that means the predicted quantile exactly attains the value of the observed value of the target variable. Overprediction is more heavily penalized than underprediciton for quantiles $\tau < 0$, while the opposite is true when $\tau > 0$. This is because the loss function places more weight on the residuals above the predicted value for overprediction and below the predicted value for underprediction.

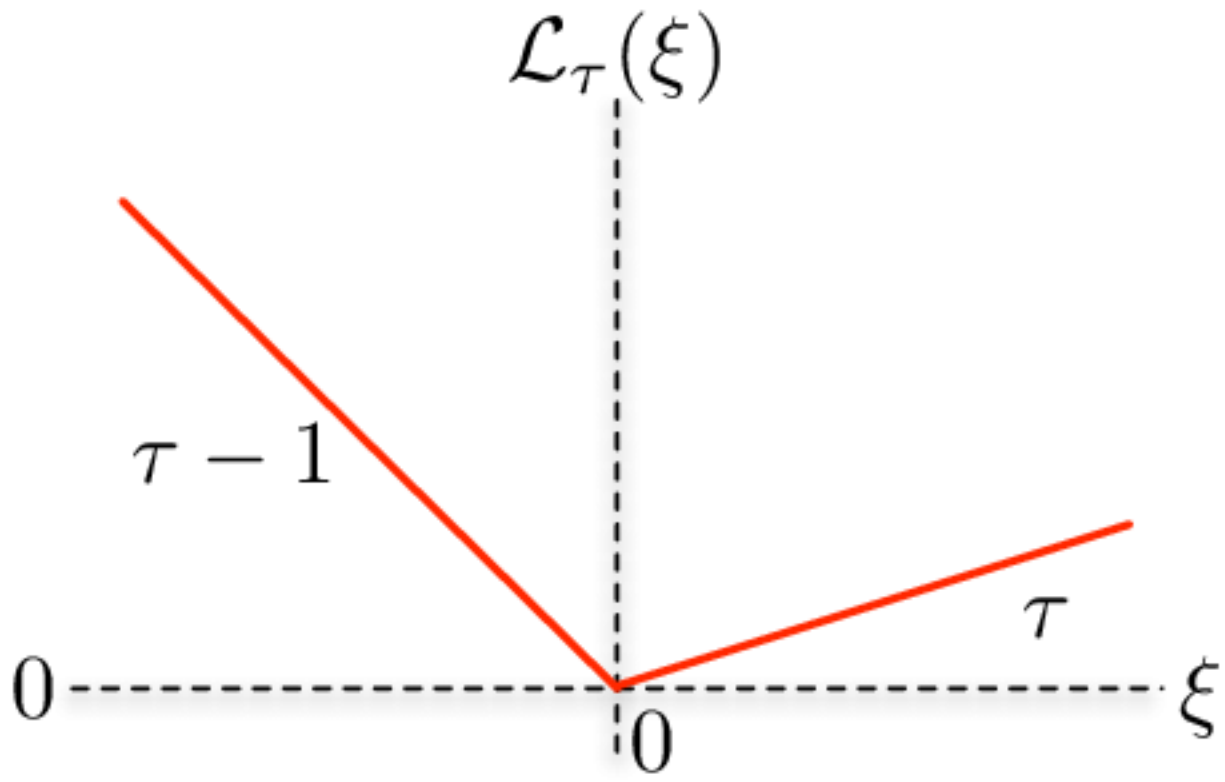

Figure 2: Illustration of the pinball loss function.

Section 3.1: Nonparametric quantile regression

By nonparametric regression models we refer to models that do not make any parametric assumptions about the form of the conditional distribution function, the relationship between the response variable and predictor variable(s) are linear however. By applying the theory of nonparametric models to a linear predictor of the conditional median, we arrive at the basis of a nonparametric quantile regression model.

For a nonparametric QR model, consider the observed response date y_i of grain i and the predictor variable year t_i with an intercept in the $(n \times 2)$-matrix X.

$$X = \begin{pmatrix} 1 & 1 & ... & 1 \\ t_1 & t_2 & ... & t_n \end{pmatrix}$$

Let Y_i be a stochastic variable corresponding to the observed value of y_i with the conditional distribution function

$$F_{Y_i|x_i}(y) = \mathbb{P}(Y_i \leq y \,|\, x_i), \quad -\infty < y < \infty \quad (?)$$

In order to find the τ-quantile of the conditional distribution of $Y_i \,|\, x_i$ we need to find the smallest value of y such that the conditional cumulative distribution function is greater or equal to τ.

So, for each quantile $0 < \tau < 1$ the inverse

$$Q(\tau \mid x_i) = \inf\{y; F_{Y_i \mid x_i}(y) \geq \tau\} \quad (?)$$

of the conditional distribution function in (?) represents the conditional quantile function.

Moving forward, our model takes the form

$$Q(\tau \mid X) = X\beta(\tau) + \varepsilon(\tau) \quad (?)$$

whereby the column vector $\beta(\tau) = (\beta_0(\tau), \beta_t(\tau))^\top$ contains the intercept $\beta_0(\tau)$ and year as the slope $\beta_t(\tau)$ of quantile τ. We define $\varepsilon(\tau)$ as a non-parameterised vector of the error terms.

An optimization problem is then constructed as to minimize the objective function in (?). In ordinary linear regression one typically uses the method of least squared errors (LSE) to find the model that leads to the lowest amount of loss of information. In our case however, we want to learn the parameters of a quantile regression model that can accurately predict the conditional quantiles of the target variable. To do this, we minimize the pinball loss function $\mathcal{L}_\tau(\xi)$, which we previously introduced in the **Methods** section, over a set of data X, y [@koenker2005qr] [@takeuchi2006nonparametric].

The resulting regression parameter estimates of the τ-quantile is thus given by the solution to the minimization problem

$$\beta(\tau) = \arg\min_{b \in \mathbb{R}^2} \sum_{i=1}^{n} \mathcal{L}_\tau(y_i - x_i b). \quad (?)$$

Comparing equation ? with equation ? (X_tau), note that we replace the scalar μ_τ with $x_i b$ in the loss function and minimize over b leading to an estimation of the β-coefficient which coupled with a vector of data x grants a predictive linear model.

Since the function (?) is non-differentiable at 0, no direct solution exists. Rather we use numerical estimation methods provided in the `R` library `quantreg` [@koenker2021rpackage], such as the Frisch-Newton interior point method to find the optimal point along the ξ-axis.

Section 3.2: Linear regression on emprical quantiles

Another approach we used is to perform linear regression on empirical quantiles, a statistical method that combines linear regression with year-wise empirical quantiles to model the relationship between a response variable and one or more predictor variables. The basic gist of EQ regression is to fit a linear regression model to the empirical quantiles of the response variable, rather than to the mean. This method is called "empirical" quantile regression because it estimates quantiles of the response variable based on the empirical distribution of the data, rather than assuming a specific distribution for the response variable.

One way to perform empirical quantile linear regression is by using an indicator function $I(\cdot)$. The indicator function is a mathematical function that takes a value as input and returns 1 if the value satisfies a certain condition and 0 otherwise. In the context of empirical quantile linear regression, the indicator function is used to define the estimation problem in terms of a set of linear programming (LP) constraints.

Rather than using the raw gregorian date y_i as response variable, we predict an empirical quantile of the dates of a subset of observations. For each year $t \in \mathcal{T}$ we extract the set of observations

$$\mathcal{Y}_t = \{y_i : I(t_i = t)\} \quad (?)$$

Let $\tau \in (0, 1)$ be a quantile. For each of our sets $\mathcal{Y}$ we let $\hat{F}_{(t)}$ be the empirical distribution function formed by the elements of its set. The corresponding empirical quantile is defined as

$$\hat{Q}_{(t)}(\tau) = \inf\left\{y \in \mathcal{Y}_t : \hat{F}_t(y) \geq \tau \mid \mathcal{Y}_t \neq \emptyset\right\}. \quad (?)$$

As in the previous method we construct the $(n \times 2)$ matrix X by stacking the intercept and vector of years $(1, t)$ on top of each other. For all $t \in \mathcal{T}$, we stack the quantiles into the vector of observations $Y(\tau)$. Next we formulate the linear model

$$Y(\tau) = X\beta(\tau) + \varepsilon(\tau). \quad (?)$$

As before $\beta = (\beta_0, \beta_t)$ defines the regression parameters for the intercept β_0 and year β_t. Recall that in the nonparametric method, no assumptions were made about the error terms. In this approach however, we assume a normal distribution of the error terms $\varepsilon(\tau) \sim N(0, \sigma^2(\tau)I_T)$, where I_T is the identity matrix of rank T.

The log-likelihood of the model is given by

$$l(\beta(\tau), \sigma^2(\tau) \mid Y(\tau), X) = \sum_{t \in \mathcal{T}} \log\, f\left(\hat{Q}_t(\tau) \mid (X\beta(\tau))_t, \sigma^2(\tau)\right) \quad (?)$$

So far in this approach the model grants each year t the same weight, not each individual pollen. Since the amount of pollen observed each year does not remain constant. A reweighting of the log-likelihood may be conducted as to grant each pollen the same weight. To get even weighting of each pollen, we add a weight factor $w_t = |\mathcal{Y}_t|$ to each empirical quantile $\hat{Q}_t(\tau)$. The reweighed log-likelihood (?), takes the form

$$l_w(\beta(\tau), \sigma^2(\tau) \mid Y(\tau), X) = \sum_{t \in \mathcal{T}} w_t \log\, f\left(\hat{Q}_t(\tau) \mid (X\beta(\tau))_t, \sigma^2(\tau)\right), \quad (?)$$

This can easily be implemented with the `weights` argument of the `lm` function where we set the weight to be $\frac{1}{N_t}$ where N_t represents the number of observations at year t.

To proceed fitting the model, we attempt to find the maximum likelihood estimate (MLE) of $\beta(\tau)$ and $\sigma^2(\tau)$ by optimizing

$$MLE(\beta(\tau), \sigma^2(\tau)) = \arg\max\, l_w(\beta(\tau), \sigma^2(\tau) \mid Y(\tau), X) \quad (?)$$

for a given quantile τ. Since the function is continuous and twice differentiable, the solution can easily be found using conventional methods.

Section 4: Results

Section 4.1: Statistical model performance across quantiles

Putting both these methods to the test, we take a look at how each method behaves for any quantile of data. In figure (?) we limit the scope of our analysis to pollen monitored in Stockholm, similar patterns are however present for data found at other stations. We observe that the intercept begins at around day 50 for both models but increases more rapidly over the quantiles for QR than with EQ, leading to a significantly later estimation of when the pollen season ends (high quantiles) using the QR approach. Another phenomenon visible in figure (?) is that the direction of the slopes are different in the second graph. What we observe is that the QR estimates of the `year` variable move from strong negative coefficients, to lesser negative or even slightly positive coefficients across the quantiles. This means we get a significantly earlier start of the pollen season, but a less shifted end to the season. All in all this results in a longer and more spread out pollen season. In stark contrast to the QR method, when analyzing the EQ coefficient estimations, we find that the early parts of the pollen season actually moved to later dates, while most of the season beyond the first 20% of pollen or so, saw a shift of around 0.3 days per year to earlier dates.

Based on this information, one may ask oneself which of these approaches we deem to be the most fit to represent the behaviour of the pollen season. The fact that the size of the confidence intervals differ so greatly is, to us, sufficient material to determine the EQ model as more desirable for the purpose of this research paper. In order to understand this difference in confidence between the to methods, let us take a look at how the models look overlayed on the set of observations.

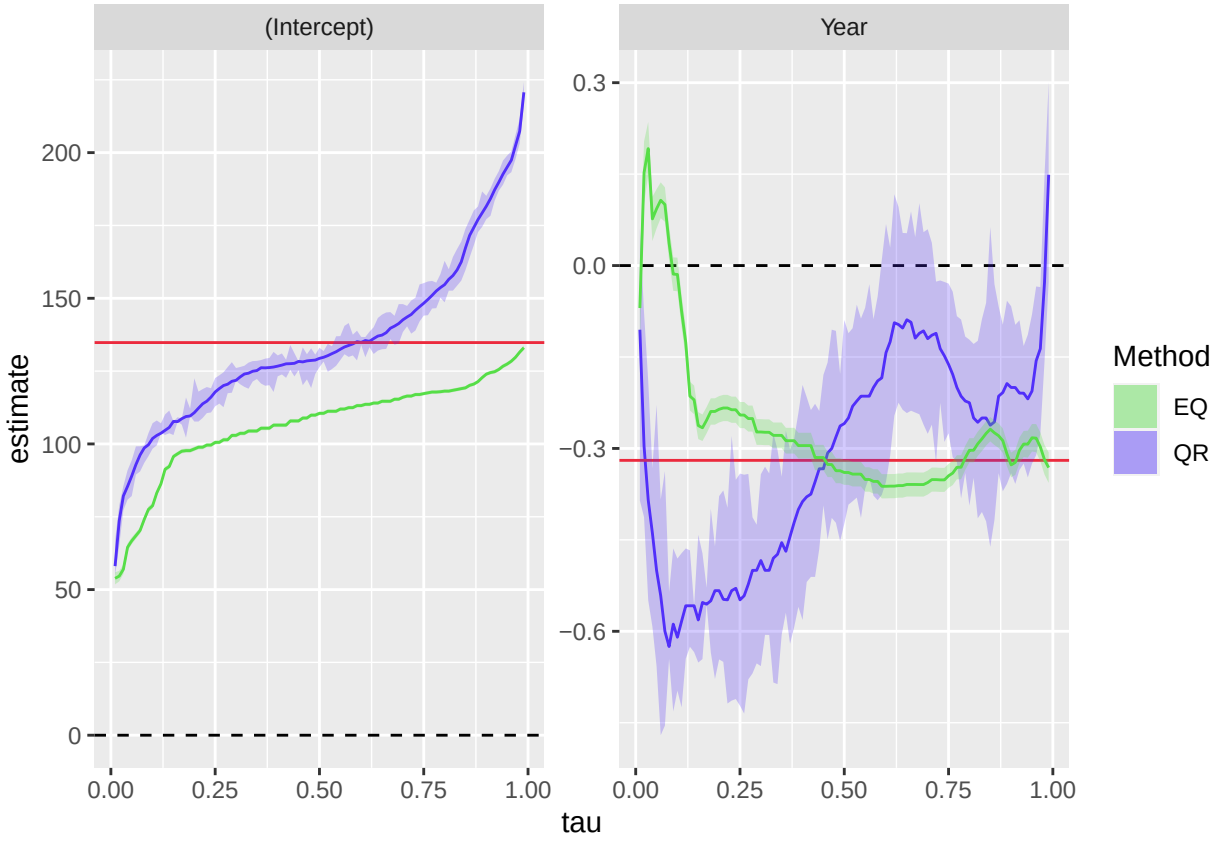

Figure 3: Line graph over the coefficient and intecept estimates for each quantile in Stockholm, all species. The thicker, opaque bounaries define the 95%-confidence intervals for each method. The red line shows the linear regression coefficient without quantile parameterization.

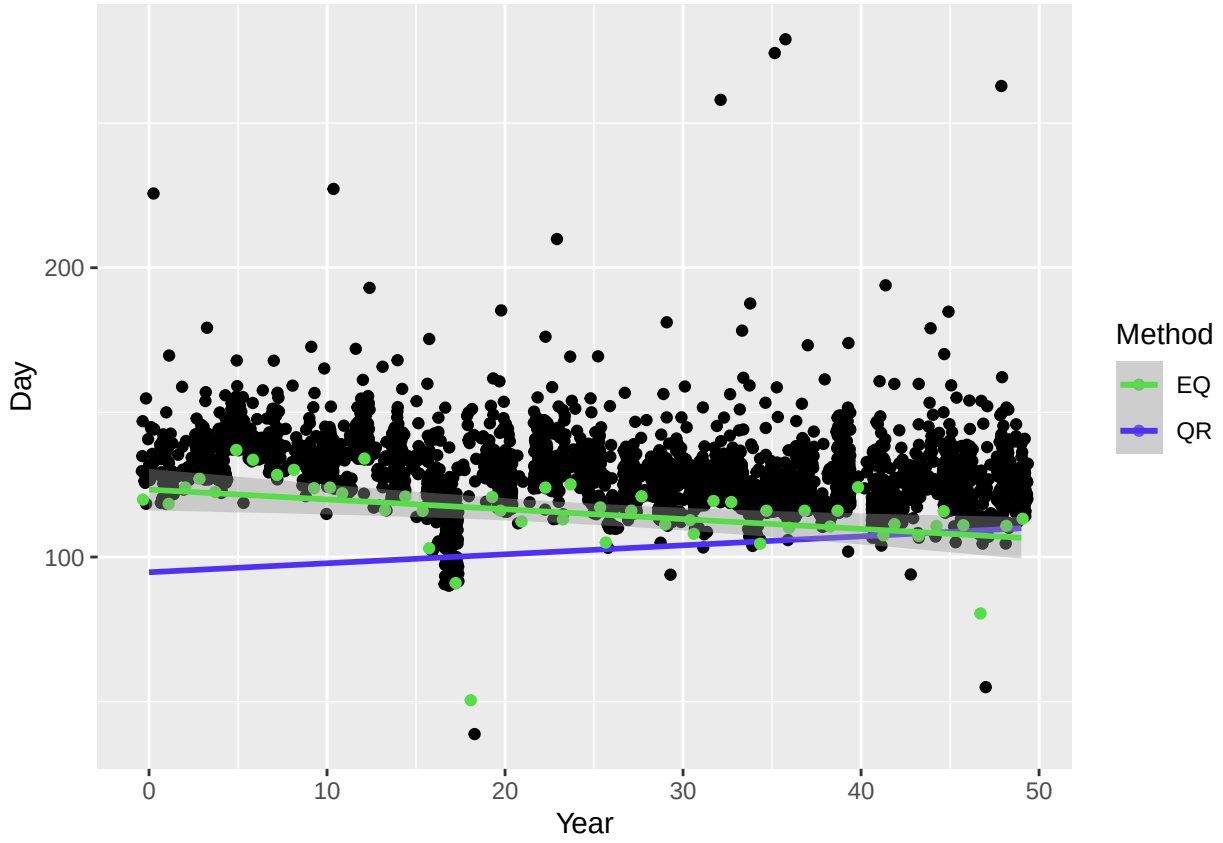

Figure 4: Jitter plot over the annual distribution of betula pollen in Stockholm comparing QR and EQ at the 1% quantile level.

Figure (?) tells us the difference in fitting of both our models. The quantile regression method fits a linear model to encompass a given quantile, in this case the first 1% of annual pollen release, below the regression line. We can observe that year 17 appears to be a heavy outlier in which the entire pollen season came significantly earlier than other years. Since the QR method tries to fit a linear trend with a specific amount of observations below it, and a large amount of these first 1% of observations are from the 17th year, the model is heavily skewed by the observations from this year and thus does not generate a good fit. EQ does not succumb to this issue since we in this approach aggregate the observations from each year to a specific date value (red points) and then perform ordinary linear regression on these aggregated observations. In this case the one outlier year corresponds to only 2% of the total observations while for QR this effect appears to be well over 50%. These results explain why the confidence intervals for QR are much larger than the equivalent measure of EQ. Thus we conclude that linear regression on empirical quantiles is a more fitting approach to use in the case of estimating the annual seasonal shift of pollen.

Section 4.2: Estimations and predictions per species

Table 3: Modelled pollen season shift by species averaged over all monitored stations.

Species	Coefficient (1%)	Coefficient (50%)	Coefficient (95%)	Estimated start 1973	Estimated start 2023	Estimated start 2050	Season length 1973	Season length 2023	Season length 2050
Corylus	-0.755	-0.539	-0.408	April 7	February 23	January 31	19	40	52
Alnus	-0.490	-0.418	-0.384	April 4	March 1	February 11	21	36	44
Salix	-0.432	-0.396	-0.178	April 20	April 1	March 22	35	50	58
Ulmus	-0.336	-0.184	-0.303	April 25	April 1	March 19	17	25	30
Quercus	-0.326	-0.217	-0.163	May 29	May 7	April 25	11	20	26
Betula	-0.188	-0.135	-0.144	April 30	April 20	April 14	25	25	26
Poaceae	-0.092	-0.201	0.132	June 1	May 28	May 26	58	72	80

Table (?) grants us information about how our EQ models predict the pollen seasons, for all monitored species, to respond to increasing temperatures in the atmosphere. The entries in the columns named `Coefficient` are the average of the β-coefficient values generated at different combinations of `species` and `station`, for each species of pollen. Although some pollen monitoring stations generated results with positive coefficients for certain species, such a phenomenon appears not to be present when we take the mean of our regression coefficients for each species from all stations. The exception to this however would be the 95%-quantile estimate for grass pollen (poaceae), meaning the grass pollen season is moving to later dates over time, while its start and peak shifts to earlier dates. This brings us to the conclusion that the pollen season unanimously is headed towards an earlier start, as well as peak, as the climate warms. The date at which 95% of all annual pollen gets monitored also seems to arrive earlier each year. The exception to this pattern is, as we point out, grass pollen. The coefficient estimates for the end of season arrival dates are however lower than the starting dates across all species. By using the linear regression coefficients shown in the table, we can predict the dates at which quantiles of the annual pollen release are expected to appear at.

Using the 1% quantile estimate, we observe that the predicted dates for hazel and alder pollen moved from early April back in 1973 to late February as of 2023 and are expected to reach late January/early February respectively by 2050. The willow and elm seasons appear to shift at a similar pace, while birch pollen has accumulated a movement of 10 days over the past half century. Oak pollen now starts to appear in early May rather than towards the end of the month which appears to have been the case in the 1970's. Just as we saw in the ridge plots in the `Annual distribution of pollen` section, the starting dates of the grass pollen season are not seen to have been greatly affected by climate change. Something we can observe across the board in regards to the monitored pollen species however, is that the season lengths (difference in dates

between the 95%-quantile and 1%-quantile) become longer over time. This phenomenon is present in all species we have covered in this paper. The largest growth in pollen season length is seen in hazel pollen (21 days in 50 years and 33 days in 77 years), while birch pollen saw the smallest growth with the season being barely a day longer after 77 years of increasing temperatures.

There could be many explanations for why the annual distribution of pollen behaves the way we have observed. In this paper we are not aiming to try and answer this question. Some quick thoughts however are that with an ever warmer climate, the temperatures at which plants are able to pollinate are present more days of the year, meaning the window of opportunity for trees and grass to release their pollen gets longer over time. The early pollen is what, for most species, gets shifted most heavily to earlier dates. An idea to why this may be the case is that since the spring arrives earlier over time, plants are able to begin the process of pollination earlier and consequently release most of their pollen (which always occurs in the spring months for arboreal species) at earlier dates, leaving the later parts of the year with less relative amounts of pollen.

Section 4.3: Geographic influence

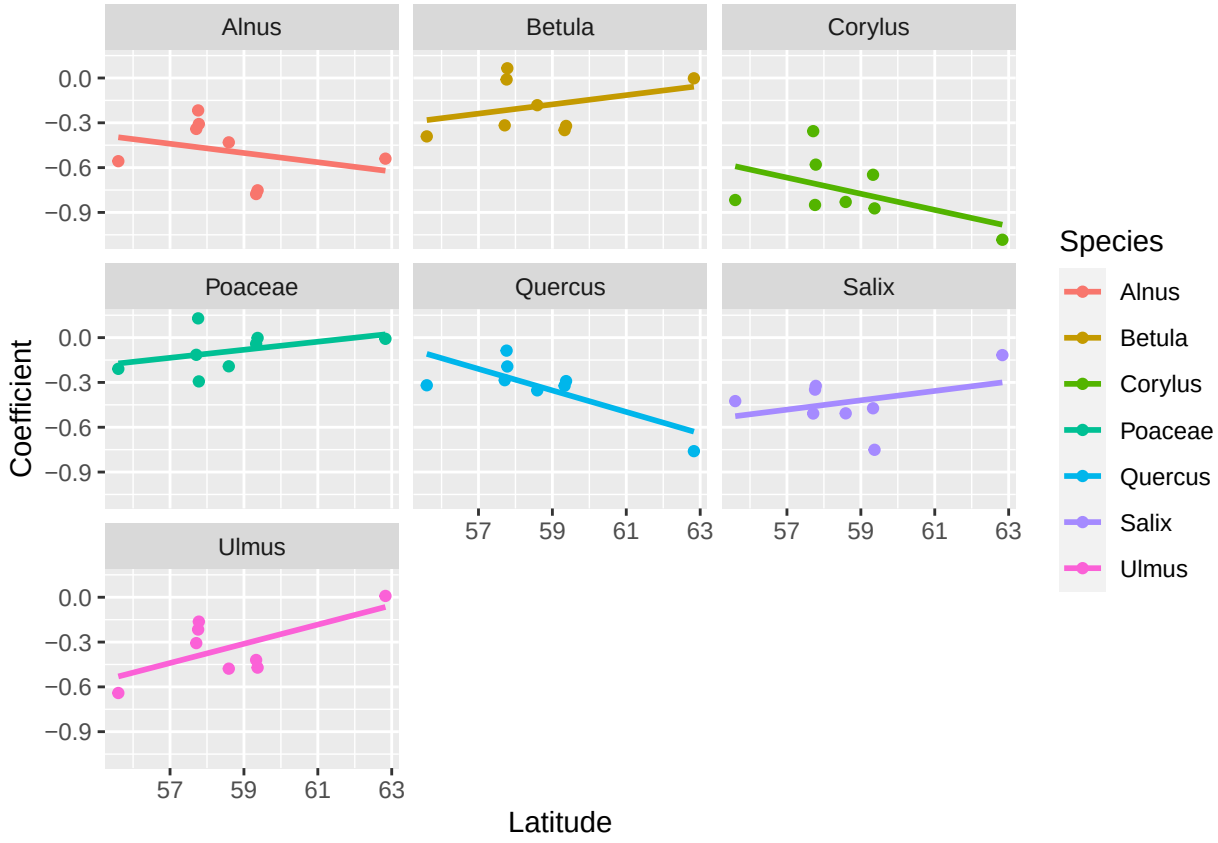

Figure 5: Visualization of latitudal impact on the EQ estimations (1% quantile)

In figures (?) through (?) we can view the regression coefficients for each species and how they differ based on the latitude of the station the pollen were collected at. Thus we can determine whether pollen in colder climates are affected by climate change by a greater extent than in warmer areas, or vice versa. Figure (?) tells us how the beginning of the pollen season shifts at the observed locations throughout Sweden. Alder, hazel and oak all appear to have more dramatic seasonal shifts at northerly latitudes, while grass, birch, willow and elm show the opposite effect. The median quantile, which we also refer to as the peak of the pollen season was not observed to have had a significant impact of the latitude of where it took place, as seen in figure (?). The steep regression slope of the peak of the elm pollen seasonal shift comes down to the fact that the coefficient at the highest latitude is vastly different to the others. By that metric, a conclusion

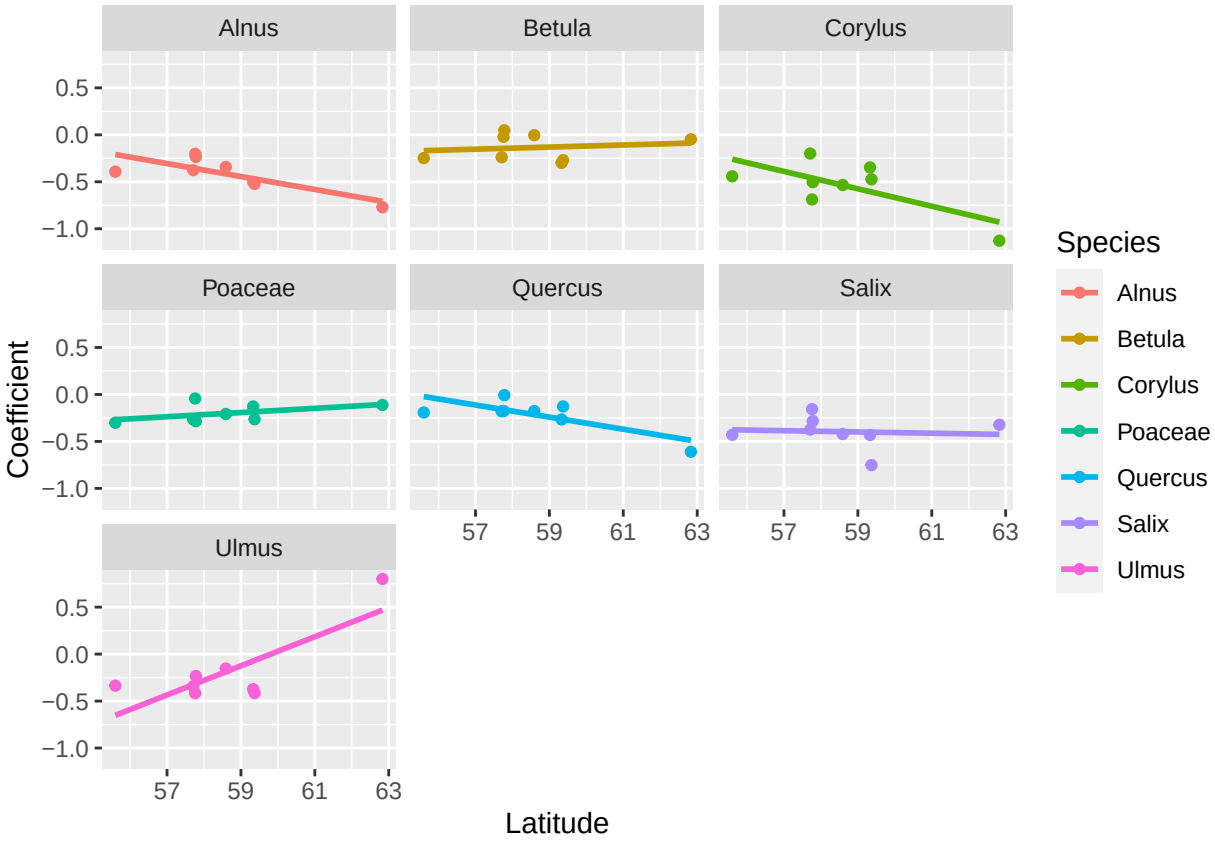

Figure 6: Visualization of latitudal impact on the EQ estimations (50% quantile)

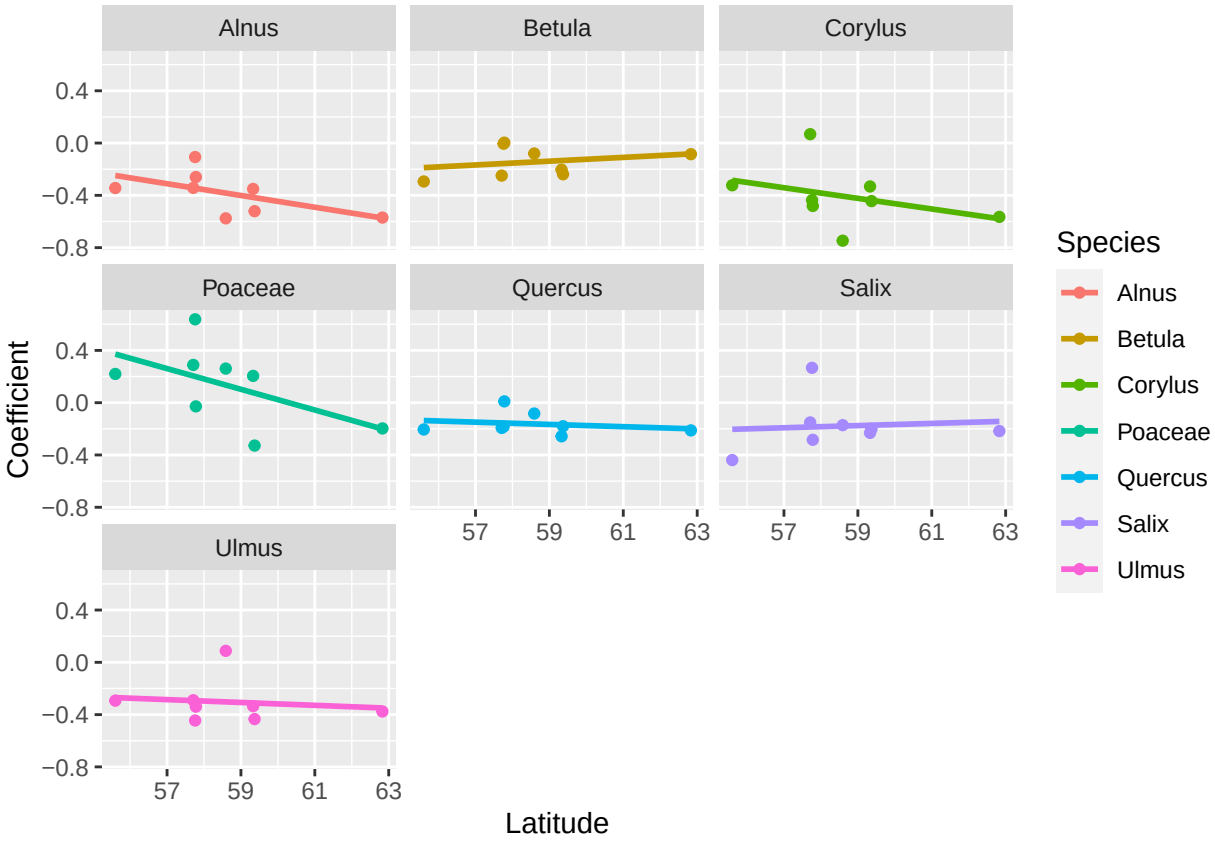

Figure 7: Visualization of latitudal impact on the EQ estimations (95% quantile)

can not be made about whether this is proof of climate change impacting elm pollen at higher latitudes simply an outlier. It is worth noting that the p-value of the aforementioned elm coefficient in Umeå is 0.981. By all accounts, this points to the data point being an outlier. By figure (?) we see how the coefficients for the 95%-quantiles are estimated. We now observe, in contrast to the 1%-quantiles, that grass pollen attributes to a quicker growth toward earlier dates, the further north one goes. The remaining pollen species' seasonal ends do not appear to be strongly affected by latitude. What we can conclude from this is that the seasonal window for grass pollen does not expand as much at high latitudes than what can be observed at lower latitudes. Alder pollen appears to, somewhat consistently, attain more strongly negative coefficients, for any quantile, at the higher latitudes. Other species were not observed to be affected by the latitude of their release by any greater extent.

Section 5: Discussion

I diskussionen ska det tillkomma text som behandlar resultaten i en mer detaljerad omfattning. Jobbar på det under veckan. Tar gärna emot feedback på vad du tycker jag ska klämma in här eller om det är något jag missat.

(Comparisson with Karlsson/Hössjer)

Section 5.1: Further improvements

The QR approach struggles with clusters of outliers such as a year with a significantly earlier pollen season. One way to look at it is that there exists a random effect of year that we do not consider in this approach. Thus deeming the assumption of independent observations obsolete in the case of our QR model. In order to improve these models, one may consider adding a random variable to deal with this yearly effect, resulting in a linear quantile mixed model.

Bibliography and References

Lägger till referenslistan när texten är klartskriven.

Appendix

Lagt in några saker jag funderar på att ha med som appendix, ska få mer ordning på det under veckan.

Appendix 1: Translation of the latin names of pollen species

Table 4: Pollen species translation in to english and swedish.

Latin name	English name	Swedish name
Alnus	Alder	Al
Betula	Birch	Björk
Corylus	Hazel	Hassel
Poaceae	Grass	Gräs
Quercus	Oak	Ek
Salix	Willow	Viden
Ulmus	Elm	Alm

Appendix 2: Results per species and location

Table 5: Start of pollen season for different species and locations, by most significant shift to earlier dates (EQ model).

Species	Location	Coefficient (EQ)	P value (EQ)	Predicted 1973 (EQ)	Predicted 2023 (EQ)	Predicted 2050 (EQ)
Corylus	Umeå	-1.083	0.003	142	74	37
Corylus	Eskilstuna	-0.873	0.000	102	52	25
Corylus	Västervik	-0.850	0.002	97	49	23
Corylus	Norrköping	-0.830	0.002	108	49	17
Corylus	Malmö	-0.817	0.000	86	45	24

Table 6: Start of pollen season for different species and locations, by most significant shift to later dates (EQ model).

Species	Location	Coefficient (EQ)	P value (EQ)	Predicted 1973 (EQ)	Predicted 2023 (EQ)	Predicted 2050 (EQ)
Poaceae	Västervik	0.129	0.323	141	150	155
Betula	Jönköping	0.065	0.575	111	114	115
Ulmus	Umeå	0.009	0.981	159	116	93
Poaceae	Eskilstuna	-0.002	0.987	149	149	149
Betula	Umeå	-0.002	0.989	132	123	118

Table 7: Peak of pollen season for different species and locations, by most significant shift to earlier dates (EQ model).

Species	Location	Coefficient (EQ)	P value (EQ)	Predicted 1973 (EQ)	Predicted 2023 (EQ)	Predicted 2050 (EQ)
Corylus	Umeå	-1.127	0.001	147	78	40
Alnus	Umeå	-0.771	0.000	130	82	56
Salix	Eskilstuna	-0.752	0.000	139	110	95
Corylus	Västervik	-0.688	0.011	99	70	54
Quercus	Umeå	-0.610	0.025	192	140	112

Table 8: Peak of pollen season for different species and locations, by most significant shift to later dates (EQ model).

Species	Location	Coefficient (EQ)	P value (EQ)	Predicted 1973 (EQ)	Predicted 2023 (EQ)	Predicted 2050 (EQ)
Ulmus	Umeå	0.800	0.108	145	137	133
Betula	Jönköping	0.048	0.660	128	122	119
Betula	Norrköping	-0.004	0.969	127	120	116
Quercus	Jönköping	-0.007	0.946	147	143	140
Betula	Västervik	-0.021	0.852	128	121	117

Table 9: End of pollen season for different species and locations, by most significant shift to earlier dates (EQ model).

Species	Location	Coefficient (EQ)	P value (EQ)	Predicted 1973 (EQ)	Predicted 2023 (EQ)	Predicted 2050 (EQ)
Corylus	Norrköping	-0.747	0.000	126	86	64
Alnus	Norrköping	-0.576	0.001	118	95	82
Alnus	Umeå	-0.570	0.000	141	98	75
Corylus	Umeå	-0.565	0.284	158	88	50
Alnus	Eskilstuna	-0.521	0.000	126	89	70

Table 10: End of pollen season for different species and locations, by most significant shift to later dates (EQ model).

Species	Location	Coefficient (EQ)	P value (EQ)	Predicted 1973 (EQ)	Predicted 2023 (EQ)	Predicted 2050 (EQ)
Poaceae	Västervik	0.638	0.005	180	231	259
Poaceae	Göteborg	0.289	0.064	210	226	235
Salix	Västervik	0.267	0.119	128	146	156
Poaceae	Norrköping	0.261	0.146	198	221	234
Poaceae	Malmö	0.220	0.162	208	221	229

Appendix 3: Information about reduced data sets (ignore)

In order to make the process of analyzing results a lot quicker, we have reduced the size of each data set to less than 5000 observations. Each data set is made up of observations of individual pollen for all combinations of species and monitoring station. Some data sets are already below this size limit, like all the hazel (corylus) data sets and a a few more at the Umeå station, while others need massive reductions in size. Figure (?) denotes sizes of the data sets for combinations of species and station. The denominators of which we reduce the data sets with are represented in brackets. As mentioned in the section **Data selection**, a reduction in the size of each data set does not affect the shape of their distribution since we use the `slice()` function to remove observations uniformly across all dates. s

Species	Eskilstuna	Göteborg	Jönköping	Malmö	Norrköping	Stockholm	Umeå	Västervik
Alnus	45 362 [10]	31 722 [7]	16 454 [4]	55 346 [12]	24 089 [5]	31 596 [7]	42 217 [9]	33 653 [7]
Betula	494 464 [99]	451 397 [91]	231 147 [47]	212 232 [43]	372 668 [75]	286 630 [58]	203 475 [41]	239 006 [48]
Corylus	3 594 [NA]	3 374 [NA]	3 812 [NA]	4 254 [NA]	113 [NA]	5 848 [NA]	3 812 [NA]	4 484 [NA]
Poaceae	50 537 [18]	31 676 [7]	74 372 [15]	48 265 [10]	37 613 [8]	22 639 [5]	52 569 [11]	48 265 [11]
Quercus	26 899 [6]	42 482 [9]	26 482 [6]	54 957 [11]	57 914 [12]	40 942 [2]	182 [2]	81 161 [17]
Salix	15 398 [4]	14 715 [3]	17 940 [4]	13 929 [3]	19 285 [4]	31 227 [7]	8 127 [2]	8 875 [2]
Ulmus	167 [NA]	33 284 [7]	6 509 [2]	35 560 [8]	4 884 [NA]	11 645 [3]	9 779 [2]	9 821 [2]