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

Max Brehmer

2023-03-21

Abstract

Preface

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]. In the same time frame Sweden has also seen significant shifts in its otherwise stable and temperate 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 7 unique species of pollen. These 7 species 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.

Prior research

Data

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. 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.

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 light description of the meaning of each variable is shown in figure ?. Data is recorded during the predicted pollen season, based on historic results.

Variable	Type	Decription
Station	categorical	Geographic location of the pollen monitoring station.
Pollen type	categorical	Genus of the recorded pollen counts.
Date	continous	Gregorian calendar date on which the airborne pollen were registered.
Count	continous	Amount of individual pollen were collected.
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.

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)

Data selection

First we remove any observations with missing values in any of the columns `date`, `station` or `name` since these are fundamental parameters to perform the data analysis on. We also need to make sure that at each station, for a certain species, pollen is continuously monitored from a given date. If the time series has a period of missing values, we simply remove any observations that occurred before the gap in the data to avoid missing values interfering with the models. Of course, using quantile based methods remove the most crucial aspects of falsely observed outliers, which missing values effectively are, we still do not want long periods of uninterrupted data to interfere with this research. In figure ? we highlight sections of the time series where missing data is present, and whether it has been removed from the dataset or not.

Theory

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 nonparametric quantile regression which we will refer to as QR. Both of these approaches estimate the response variable as the conditional median or other quantiles of the predictor variables as opposed to ordinary linear regression which uses the method of ordinary least squares (OLS) to estimate the conditional mean [Wikipedia].

OLS regression is the preferred method for many research purposes due to its inherent simplicity. However for it to be effective, a lot of assumptions need to be made about the data. More specifically homoscedasticity and normally distributed observations are two of the most central aspects of linear regression, meaning a linear regression model performs 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 one can make more accurate predictions about 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. 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.

Nonparametric quantile regression

By nonparametric regression models we refer to models that do not make any assumptions about the functional form of the relationship between the response variable and predictor variable(s). The advantage of this approach is that one can capture complex nonlinear relationships and is more robust when it comes to dealing with significant outliers in the data. 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 univariate 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 we use the pinball loss function $l_\tau(\xi)$ introduced in [koenker2005qr] as a metric to determine the accuracy of a quantile estimate [takeuchi2006nonparametric]. The pinball loss function is given by

$$l_\tau(\xi) = \begin{cases} \tau\xi & \text{if } \xi \geq 0 \\ (\tau - 1)\xi & \text{if } \xi < 0 \end{cases} \quad (?)$$

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

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

Since this function is non-differentiable at 0, we cannot use... Rather we use numerical estimation methods provided in the `R` library `quantreg` [koenker2021rpackage] to find the optimal point along the ξ-axis.

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 quantile regression 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 $t \in \mathcal{T}$ we extract the set of observations

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

As in the previous method we construct the $(n \times 2)$ matrix X by stacking each of the covariates $(1, t)$ on top of each other. Let $\tau \in (0, 1)$ be a quantile. For each of our sets $\mathcal{Y}$ we let $\hat{F}_{(x,t)}$ be the empirical distribution function formed by the elements of its set. The corresponding empirical quantile is defined as

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

For all $t \in \mathcal{T}$, we stack these 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_1. 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 (omitting τ) of the model is given by

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

This model grants each year t the same weight, not each individual pollen. To get even weighting of each pollen, we simply add a weight factor $w_t = |\mathcal{Y}_t|$ to each empirical quantile $\hat{Q}_t(\tau)$. The reweighed log-likelihood (?), again omitting τ from the notation, takes the form

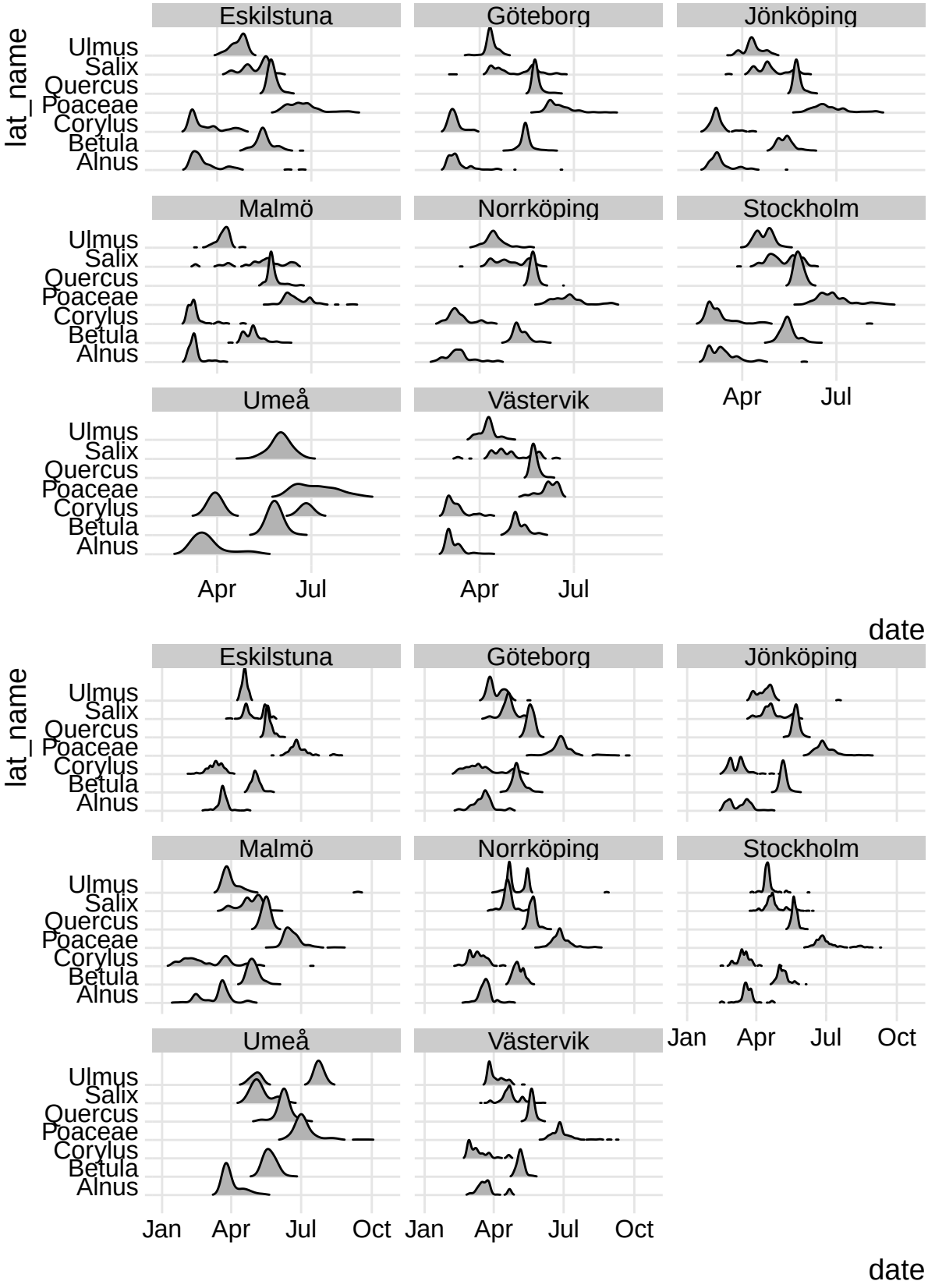
$$l_w(\beta, \sigma^2 \mid Y, X) = \sum_{t \in \mathcal{T}} w_t \log\, f\left(\hat{Q}_t(\tau) \mid (X\beta)_t, \sigma^2\right), \quad (?)$$

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 provided in the `quantreg` package.

Results


Eskilstuna
Göteborg
Jönköping
Malmö
Norrköping
Stockholm
Umeå
Västervik
lat_name
Ulmus
Salix
Quercus
Poaceae
Corylus
Betula
Alnus
Jan
Apr
Jul
Oct
date

Discussion

Further improvements

The solution to this inconvenience is to multiply the pollen count by a factor related to the size of the microscope.

To avoid presenting outlier seasons in the ridge plot, we can use an average of perhaps the first X years compared to the last X years.

Bibliography and References

Appendix

English translation of the latin names of pollen species

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

```
## # A tibble: 3,812,446 x 9
##    date       station    swe_name lat_name factor  year latitude greg_~1 md_date
##    <date>     <chr>      <chr>    <chr>     <dbl> <dbl>    <dbl>   <dbl> <chr>
##  1 1976-04-03 Eskilstuna Al       Alnus       1.2  1976     59.4      94 04-03
##  2 1976-04-04 Eskilstuna Al       Alnus       1.2  1976     59.4      95 04-04
##  3 1976-04-06 Eskilstuna Al       Alnus       1.2  1976     59.4      97 04-06
##  4 1976-04-06 Eskilstuna Al       Alnus       1.2  1976     59.4      97 04-06
##  5 1976-04-06 Eskilstuna Al       Alnus       1.2  1976     59.4      97 04-06
##  6 1976-04-07 Eskilstuna Al       Alnus       1.2  1976     59.4      98 04-07
##  7 1976-04-07 Eskilstuna Al       Alnus       1.2  1976     59.4      98 04-07
##  8 1976-04-07 Eskilstuna Al       Alnus       1.2  1976     59.4      98 04-07
##  9 1976-04-07 Eskilstuna Al       Alnus       1.2  1976     59.4      98 04-07
## 10 1976-04-07 Eskilstuna Al       Alnus       1.2  1976     59.4      98 04-07
## # ... with 3,812,436 more rows, and abbreviated variable name 1: greg_day
```