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

Max Brehmer

2023-02-19

```
## [1] "Alnus"   "Betula"  "Corylus" "Poaceae" "Quercus" "Salix"   "Ulmus"
```

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 0.2C per decade since the mid seventies, this constitutes an almost 1C 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 multiple quantile based methods, including empirical quantiles and non-parametric QR. As a result, this paper covers the majority of the theory in regards to the construction of the statistical models.

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

Methods and material

Collection of pollen data

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, for which a descriptive presentation follows in the `Data` section, this phenomenon has been ignored.

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]. 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.

Station	Latitude	Pollen_types
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)

Statistical 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 complex powerful non-parametric 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 least squares to estimate the conditional mean [Wikipedia].

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. The predictor variables to consider are expressed in a covariate vector $x_i = (1,\, t_i,\, x_{i1}, ...,\, x_{ij})$ where the discrete variable $t_i \in \mathcal{T} = \{1, ..., T\}$ represents the Gregorian year of observation. T being the amount of years monitored. The remaining j covariates are either binary dummy variables as to represent the array of different pollen genus available in the data defined by $x_{i1}, ..., x_{ij} \in \mathcal{X}_1, ..., \mathcal{X}_j$, see [Sundberg2021kompendium] for more theory about dummy variables. Moreover continous variables may also be added to complement the single parameter models by adding factors such as latitude.

Univariate models For a univariate non-parametric quantile regression model, consider the observed response date y_i of grain i and the covariates year t_i and 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

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

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

Moving forward, our model takes the form

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

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 $(?)$. [koenker2005qr] introduces the following pinball loss function as a metric to determine the accuracy of a quantile estimate.

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

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

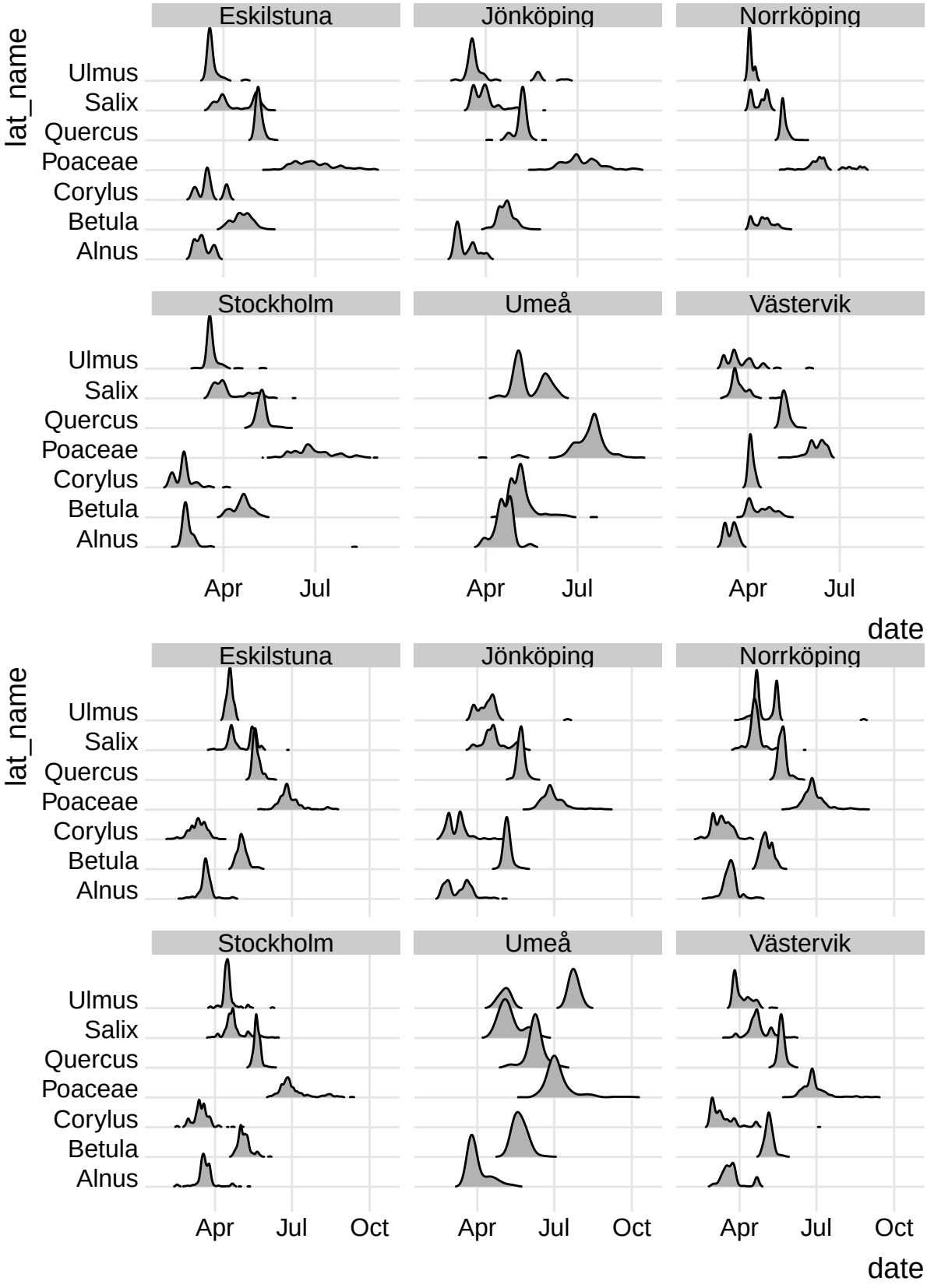
$$\beta(\tau) = \arg\min_{b \in \mathbb{R}^2} \sum_{i=1}^{n} l_\tau(y_i - x_i b).$$

To find the optimal point along the ξ-axis, we use numerical estimation methods provided in the `quantreg` `R` library [koenker2021rpackage].

Multivariate models

Non-parametric quantile regression By non-parametric regression models we refer to models that do not make any assumptions about the functional form of the relationship between the reponse variable and predictor variable(s). The advantage of this approach is that one can capture complex non-linear relationships and is more robust when it comes to dealing with significant outliers in the data. By applying the theory of non-parametric models to a linear predictor of the conditional median, we arrive at the basis of a non-parametric quantile regression model.

Results


Eskilstuna
Jönköping
Norrköping
Stockholm
Umeå
Västervik
lat_name
Ulmus
Salix
Quercus
Poaceae
Corylus
Betula
Alnus
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