Testing bayesian techniques and quantile regression to identify limiting responses of tree species

Felix Klug, Thomas Welchowski

Project partner: Karl Mellert
Supervisor: Prof. Dr. Helmut Küchenhoff

Statistical Consulting
Institute for Statistics
Ludwig-Maximilians University Munich

07.11.2013

1.Aims

Aims

- Fitting a limiting response curve for dichotome variables
- Usage of different approaches including bayesian models and quantile regression
- Comparing these models in prediction and plausibility

2. Methods

Source of Data

- Occurrence data stem from the "International Co-operative programme on assessment and monitoring of air pollution effects on forests" (ICP Forests)
- Absence at Level I monitoring plots were converted to presence if a presence has to be expected due to expert knowledge (Bohn map). [3]
- The dataset contains 7573 observations of 69 variables. 36 of these are response variables, indicating growth of different kind of trees in different regions
- All climate variables are taken from "WorldClim" a project which measures different variables with a set of global climate layers (climate grids) with a spatial resolution of 1 square kilometer [4]

Why did we try bayesian inference?

Bayesian inference has some advantages to maximum likelihood estimation:

- Interpretation of parameter and credibility intervalls
- Paramter estimations are more robust than maximum likelihood
- Prior knowdledge can be modelled

First model approach:bayesQR I

bayesQR [1] is an alternative model approach to normal logistic regression. It uses a latent variable to predict probabilitys of dichtomic variables:

$$y_i^* = x_i^T \beta + \mu_i$$
$$y_i = 1 \text{ if } y_i^* \geq 0$$
$$y_i = 0 \text{ otherwise}$$

Advantages

- Logistic regression via bayesian approach
- Can model linear separable data

First model approach:bayesQR II

Disadvantages

- High computational costs
- Parameter did not converge
- Model function is unstable

The model is available in `R` as package `bayesQR` [2]

Second model approach:INLA I

Markov Chain Monte-Carlo methods are often needed to evaluate posterior distributions. However these methods have high computional costs. To compensate these costs an alternative approach is given via the **I**ntegrated **N**ested **L**aplace **A**pproximations [5].
The Approach defines a new class of models: The "latent gaussian" models. In these models the posterior is approximated via the nested Laplace or simplified Laplace approach:

$$
\begin{aligned}
\pi(x, \vartheta | y) &\propto \pi(\vartheta)\pi(x|\vartheta) \prod_{i \in I} \pi(y_i | x_i, \vartheta) \\
&\propto \pi(\vartheta)|Q(\vartheta)|^{\frac{1}{2}} \exp\left[-\frac{1}{2} x^T Q(\vartheta) x + \sum_{i \in I} \log\{\pi(y_i|x_i, \vartheta)\}\right]
\end{aligned}
$$

INLA is available in `R` as package `INLA` [6]

Ad hoc solution I

1. The expectation $\mathrm{E}(\mathbf{Y}|\mathbf{X}) = \mathrm{h}(\boldsymbol{\eta})$ is modelled
2. h is a known distribution function
3. The quantile $Q_\tau(\boldsymbol{\eta}|\mathbf{X})$ is fitted, where τ is an extreme low or high quantile

Examples

- Alternatives for Point 1: GLM,GAM,Boosting, Boosted Trees, Feed forward neural network
- Alternatives for Point 2: Quantile regression, Additive quantile regression, Quantile regression forest, Expectile regression, Quantile regression neutral networks

Example

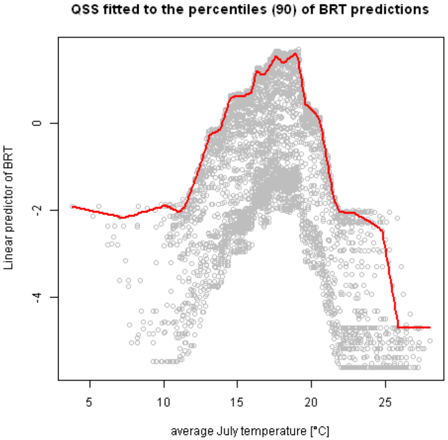
- We used a GLM for Point 1 and Quantile regression for Point 2

Ad hoc solution II

- The parameters for both approaches were exactly the same
- The quantile regression only finds the linear coherence from the GLM

Example II Boosted Trees and additive quantile regression

Ad hoc solution III



3. Results

Summary

Usuage of different model types. To this point none of our model approaches were sufficient for the problem.

- bayesQR: First approach; Model was unstable and had high computional costs
- INLA: Faster fit than bayesQR; no real quantiles; maxima and minima and the edge of the plots
- Ad hoc solution: best approach so far; no usuage of simple methods

Discussion

Do you have improvements or critical comments for the ad hoc solution? Which models would you use for the first and second point for the ad hoc approach?

Bibliography I

D. F. Benoit and D. Van Den Poel.
Binary quantile regression: A bayesian approach based on the asymmetric laplace density.
Working Papers of Faculty of Economics and Business Administration, Ghent University, Belgium 10/662, Ghent University, Faculty of Economics and Business Administration, August 2010.

Dries F. Benoit, Rahim Al-Hamzawi, Keming Yu, and Dirk Van den Poel.
bayesQR: Bayesian quantile regression, 2013.
R package version 2.1.

Udo Bohn.
Karte der natürlichen Vegetation Europas. Maßstab 1:2500000 - Map of the natural vegetation of Europe. Scale 1:2500000.
Bundesamt für Naturschutz, Bonn, Bad Godesberg, 2005.

Bibliography II

Cameron S. E. Parra J. L. Jones P. G. Hijmans, R. J. and A. Jarvis.
Very high resolution interpolated climate surfaces for global land areas.
International Journal of Climatology, 25:1965–1978, 2005.

Havard Rue, Sara Martino, and Nicolas Chopin.
Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations.
Journal of the Royal Statistical Society Series B, 71(2):319–392, 2009.

Havard Rue, Sara Martino, Finn Lindgren, Daniel Simpson, and Andrea Riebler.
INLA: Functions which allow to perform full Bayesian analysis of latent Gaussian models using Integrated Nested Laplace Approximaxion, 2013 +0200).
R package version 0.0-1379661604.

Graphics to BayesQR I

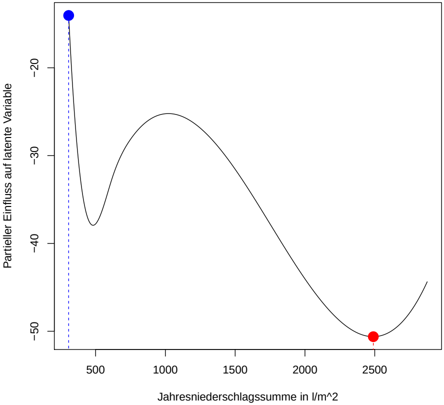


Figure : BayesQR:Partial influence plot of annual precipitation of 0.1 % quantile (Tree=Common Spruce)

Graphics to BayesQR II

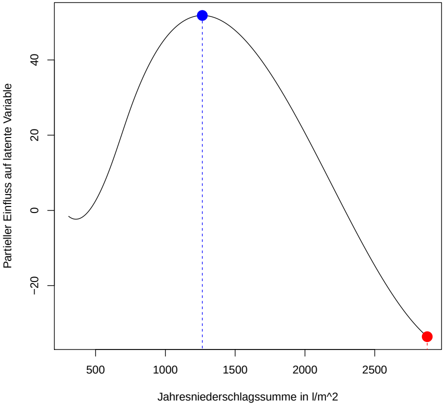


Figure : BayesQR:Partial influence plot of annual precipitation of 0.9 % quantile (Tree=Common Spruce)

Ad hoc solution II

1. First a gam model is fitted including temperature, precipitation and a spatial effect:

$$\frac{\log(P(Y_i = 1))}{1 - \log(P(Y_i = 1))} = \underbrace{\beta_0 + f(x_{i1}) + f(x_{i2}) + f(x_{i3}, x_{i4}) + \epsilon_i}_{\eta}$$

$$p(\eta) = \frac{\exp(\eta)}{1 + \exp(\eta)}$$

2. Then the partial influence of $f(x_{i3}, x_{i4})$ is predicted. The 95 % empirical quantile is fitted. All oberservations, which have predictions beneath this quantile are sorted out.
3. Now the model is refitted leaving out the spatial effect, so that only the trees, which are on the upper level of the population are fitted.

Graphics to solution II

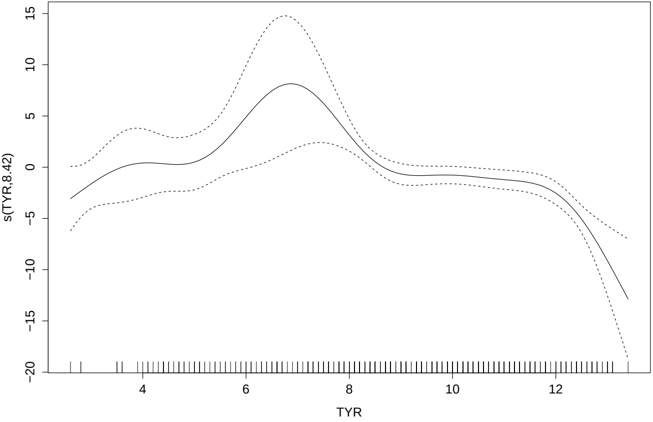


Figure : Plot to variable temperature of solution II

Graphics to solution II

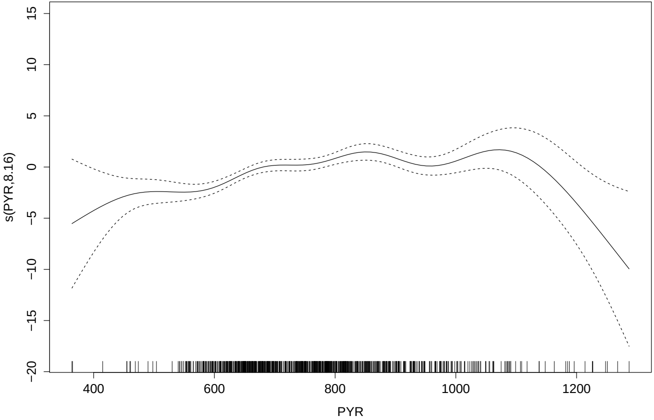


Figure : Plot to variable precipitation of solution II

INLA: Example Graphic

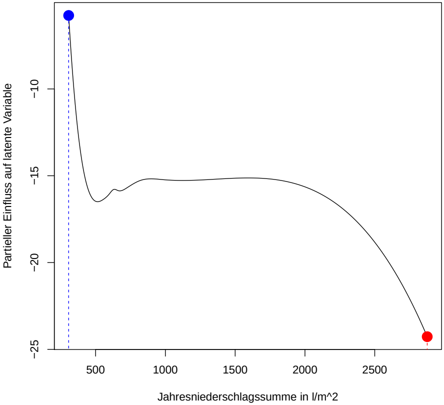


Figure : Partial influence plot of annual precipitation of 0.05 % quantile (Tree=Common Spruce)

INLA: Example Graphic

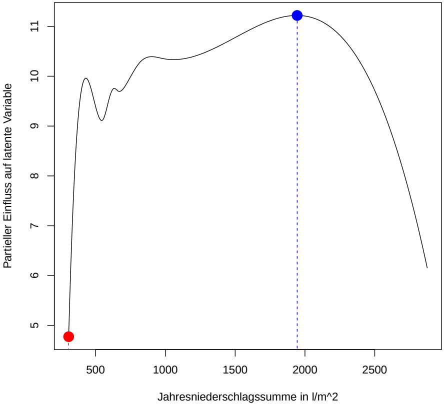


Figure : Partial influence plot of annual precipitation of 0.95 % quantile (Tree=Common Spruce)