Department of Mathematical Statistics
University of Lund, Sweden

Bachelor Thesis Fall 2012

Jesse Burström
jesse.burstrom@gmail.com

Bud burst of birch in Finland and the United Kingdom
Logistic regression analysis and modeling

Submission Date: January 23th 2013

Advisor: Johan Lindström

Abstract

The day of bud burst (DBB) of different tree species are known to be affected by factors such as growing degree days and temperature. In this paper a two state Markov chain is used to model DBB for birch. The model is fit using logistic regression and LASSO regularization is used to evaluate which of many potential factors best forecast DBB. Data of birch from both Finland and the United Kingdom is studied and differences between the models adapted to the two countries are investigated. For modeling purposes to capture the environment of forecasting, estimated interpolated gridded climate data was used and not directly measured climate data.

It is found that the models give very accurate predictions on the DBB. For Finland it is little more than 2 days in mean absolute error (MAE). The model is also fairly compact having less than 10 explaining covariates. The covariate, accumulated growing degree days, was as expected part of the models as well as among others variation of precipitation.

Acknowledgments

I want to thank Johan Lindström for giving me this interesting project to develop my R-programming skills, and understanding of logistic regression. I want to thank for all the assisting and helping writing this report. It was a journey that I could not foresee and it wouldn't have been possible without all the helpful remarks, insights and tips. I further want to thank Anna Maria Jönsson and Cecilia Olsson for bringing valuable insight into the day of bud burst for trees a branch of geographic phenology.

Climate data comes from E-OBS dataset from the EU-FP6 project ENSEMBLES (http://ensembles-eu.metoffice.com) and the data providers in the ECA&D project (http://www.ecad.eu)"

Phenology (Bud burst) data was provided by the members of the PEP725 project and obtained from: http://www.pep725.eu/

Contents

1. Introduction

This analysis sets out to investigate which climate factors best predict the day of bud burst (DBB) for birch (lat. *Betula pendula*) trees in Finland and the United Kingdom. Observations (location and year) of DBB are given together with climate covariates; temperature, precipitation, day length, latitude and elevation are considered in different combinations. This analysis is inspired by Song (2010) who analyzed the DBB of different tree species in Canada with respect to only growing degree days (GDD). This analysis expands on the work of Song by investigating the effects of many factors, and how they best make up a forecast model for predicting the DBB.

A tree goes through a series of phenological stages such as budburst and shedding of leaves. Kramer (1994) refers to Vegis (1964) three stage rest model in which only at the third stage post-rest growth is possible. The state transition between the three stages (pre-rest, true-rest, post-rest) is triggered by chilling attaining certain threshold values. See also the thermal time model by Cannell & Smith (1983). Other models propose amount of daylight as a key triggering factor for budburst. A long photoperiod could substitute for lack of chilling (Vegis, 1964). Thus day length is considered as a covariate in the many factor model. Further the amount of rain during the year before could affect budburst. Stress due to drought could affect the rest needed. Therefore different periods are also considered as possible covariates, the amount of rain along with number of rain days. The different, so called, constant covariates (see appendix for list of covariates) are considered to assess the impact of various weather conditions over the previous year and their influence on the condition of the tree. Unfavorable weather conditions could potentially delay the budburst the next year.

The aim of this analysis is to take a statistical approach on modeling the bud burst. No effort (more than the choice of the covariates made by my advisor Johan, Anna and Cecilia) is made to link actual biological relevance of the different covariates and interactions. The focus is entirely on the statistical modeling and the predictive power of the model. For previous models it has been common to look at mechanistic factors such as GDD, chilling followed by forcing temperature and photoperiod. The purely statistical approach taken here could help in finding other relevant biological factors linked to the DBB and even so could be used with climate models for prediction.

2. Data

Phenology data for trees are collected irregularly across countries and are located differently from climate data. Many phases of the tree cycle are recorded and in this model only the DBB is used. The exact definition of the DBB used is event no 11 in the BBCH scale (Feller, 1995). Event 11 is when the first true leaf has unfolded. The DBB data where retrieved by the PEP725 project (2012). Birch (lat. *Betula Pendula*) DBB data are used in this analysis.

The data for parameter estimation in the Finland model is chosen as 2/3 of the DBB observations and are randomly selected. In the UK the 418 gridded climate data locations are randomly selected and grouped hierarchal in three models in order to study the dependence on the number of observations and model resulting model size. The third model is a subset of the second being a subset of the first. The DBB stations associated with these climate locations are used for parameter estimation. The three models have 33%, 17% and 11% percent of the gridded climate data locations associated with them.

Temperature & Precipitation
Climate data were obtained from Heylock (2008) on a gridded 0,5 x 0,5 degree grid. DBB is assigned to the closest grid cell. Coastal data are dropped due to lack of climate data. Temperature and precipitation data is collected from the closest climate cell relative the DBB station. The climate data consists of daily average temperatures and minimum respectively maximum daily temperatures, as well as precipitation (amount of rainfall each day).

Day length
The day length data, or number of hours of light, is calculated from the time of year and latitude for each station.

2.1 Stations

Finland
For Finland there are 33 DBB phenological observation sites well spread in space with DBB data for birch over the years 1997-2005. Se figure 2a & b

United Kingdom
There are 3169 phenological observation sites across the UK with DBB data for birch during the years 1972-2005, with almost all of the data coming from 1999-2005. The 3169 DBB stations are situated within 418 - 0,5 x 0,5 degree cells providing climate data. Therefore many DBB stations have the same climate data. This will affect the predictions since the coefficient estimation for the different covariates have to weight in several ‘right’ answers to the same data. In effect an ‘average’ is calculated to best fit all data. Se figure 2c & d

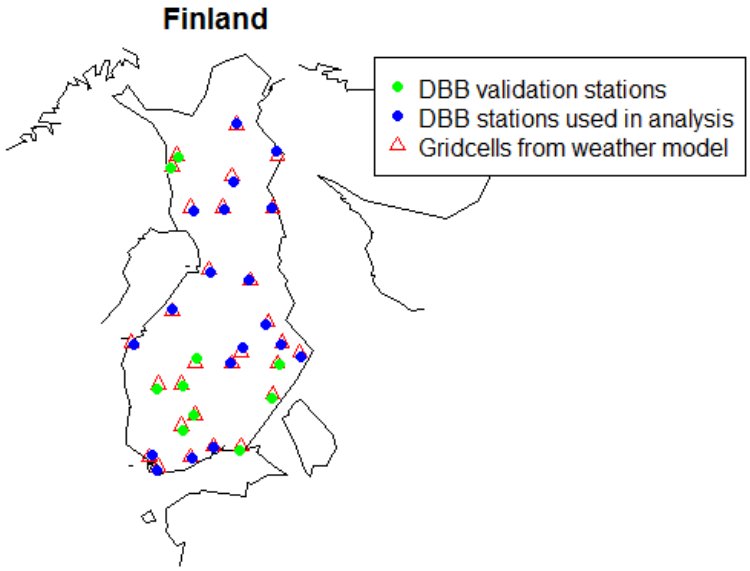

Figure 2a DBB phenological observation sites in Finland used in the analysis and corresponding gridded climate data locations closest to each DBB cell.

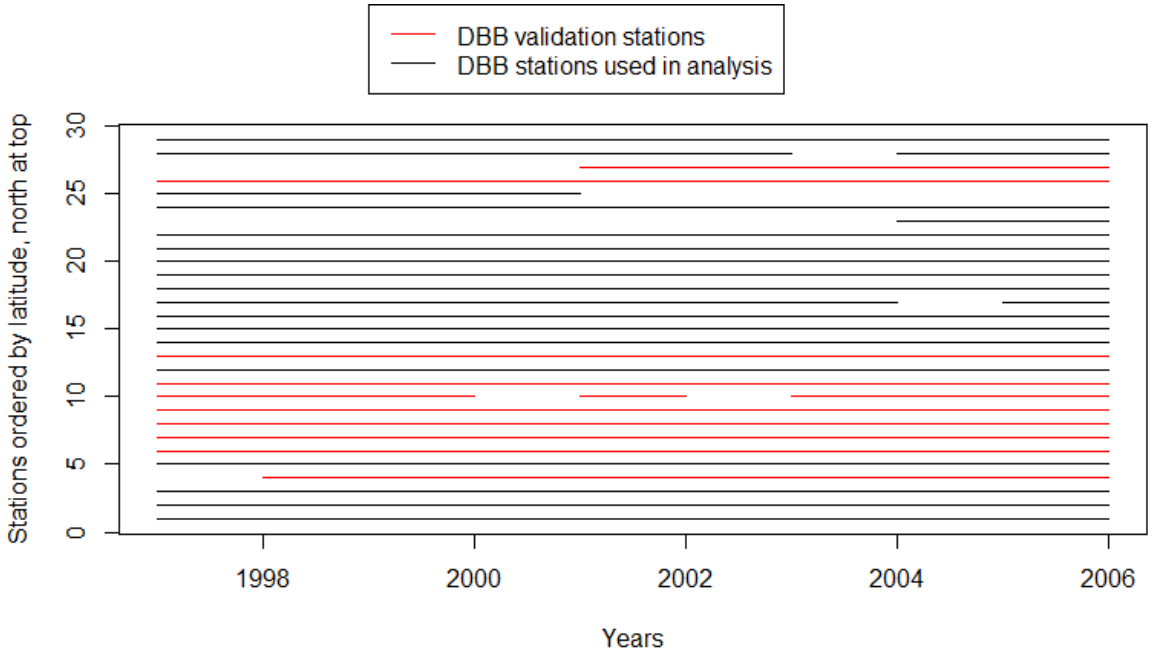

Figure 2b Availible DBB data for each station and year

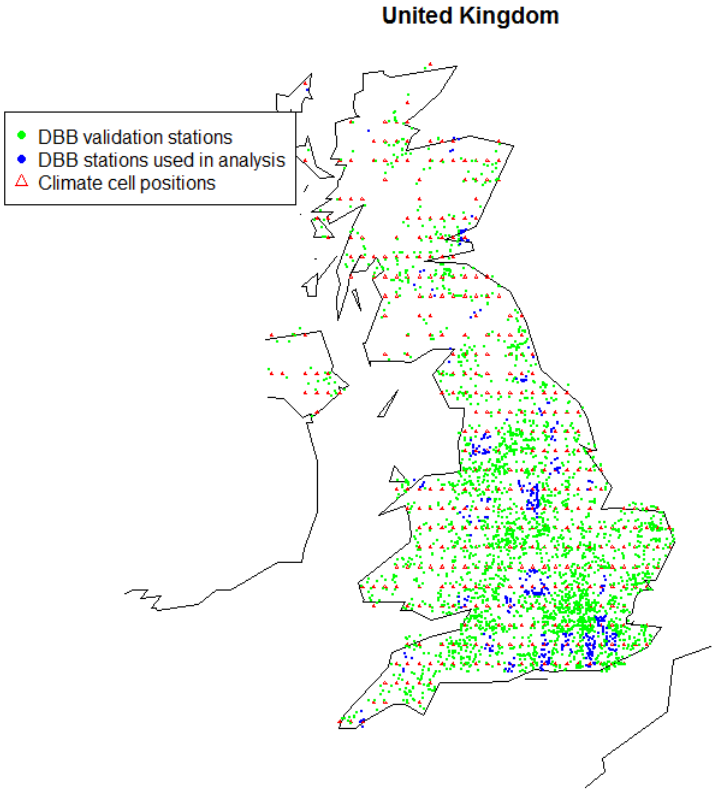

Figure 2c DBB stations in the United Kingdom used in analysis for the UK3 model having approximately 11% of the data and corresponding climate grid cells closest to each DBB cell.

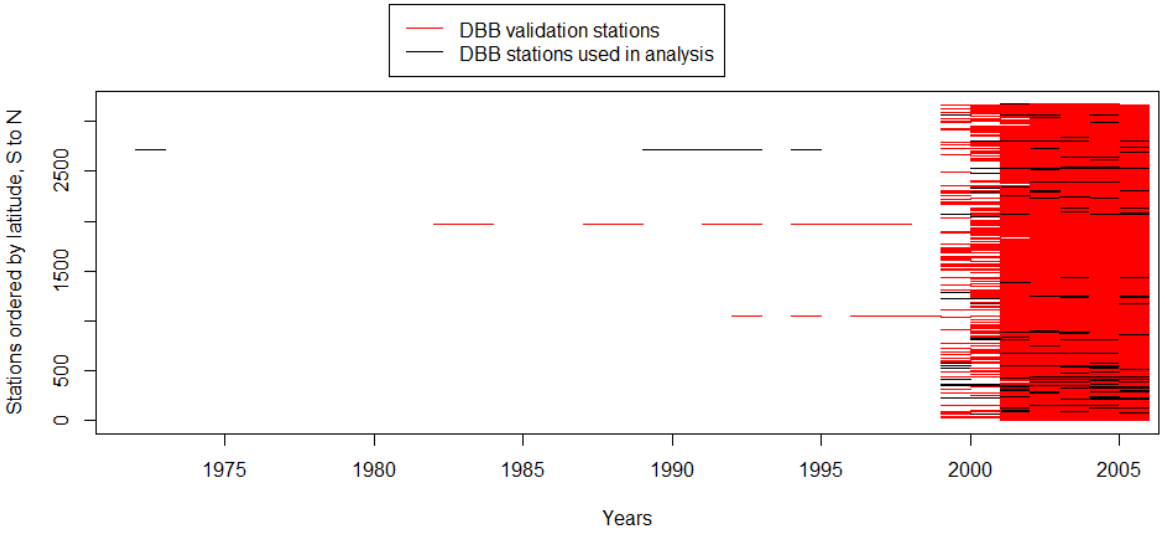

Figure 2d Used DBB data for each station and year for the UK3 model.

Table 2e Data table for Finland and the UK. Note that extreme values are kept.

	Years	Number of DBB stations	Number of climate stations	Median DBB	min/max DBB
FI	1997-05	33	29	138	119/172
UK	(1972-98) 1999-05	3169	418	103	32/150

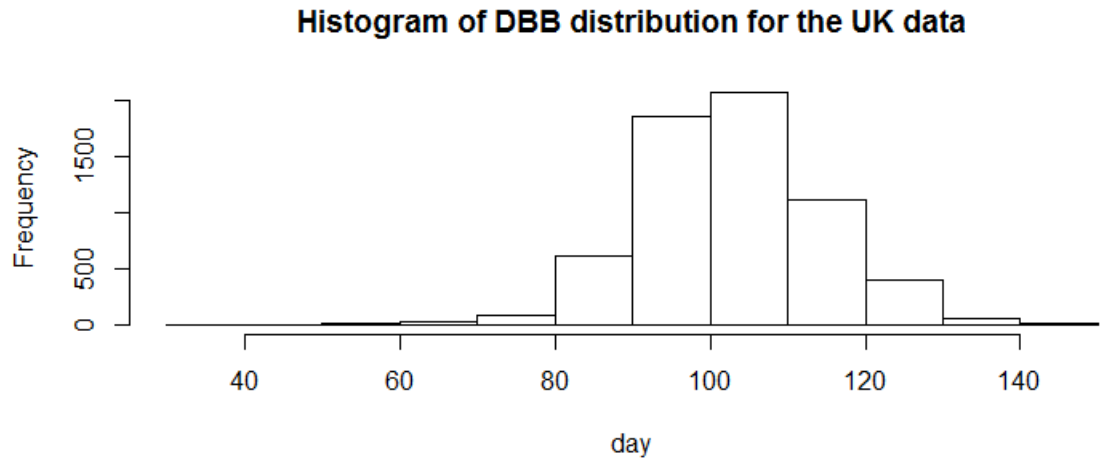

Figure 2f The distribution of the various DBB instances for the UK data. The data varies as much as from day 32 to day 150. In addition the local variation is large.

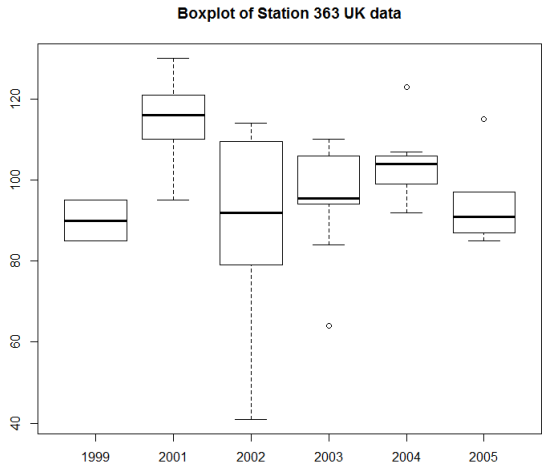

Figure 2g Boxplot of weather station 363 UK data showing the variability within a year. Number of observation sites each year in boxplot

year	1999	2000	2001	2002	2003	2004	2005
instances	2	0	10	11	14	10	5

3. Model setup

Our model focus is on the DBB, which can be interpreted as *time-to-event* data. By time-to-event is meant a sequential time dependent process having a transition at a specific time (Song, 2010, ch 2.2). The transition is an event or state change. More specifically for bud burst this transition will be absorbing from no leafs to presence of leafs. Absorption implies that when the state is reached the process cannot reverse. In a natural way a more general sequence of state transitions can be modeled. We assume that the states are 0='no leafs' from start of the year up to the DBB where it changes to 1='leaves' and then stays in that state for the remainder of the year. So for each observation of one DBB-every indexed i, we have time points $t = 0, 1, \ldots,$ defined for each day starting at January 1 and we have an indicator variable $Y_{i,t} \in \{0,1\}$ reflecting the status of the tree where 0 stands for "bud burst has not occurred" and 1 for "bud burst has occurred". Let T_i be the time to event for observation i then:

$$Y_{i,t} = 0, \; t < T_i$$
$$Y_{i,t} = 1, \; t \geq T_i$$

Now we further assume there are time dependent vectors of covariates, $X_{i,t}$,which affect T_i . In the continuing we will assume that both covariates and response, exists for all observation couples (location and year).

3.1 Equivalence between Markov chain transition model and logistic regression

We are now interested in the conditional event of DBB given a set of covariate values. Extending the notation above to cover sets of time points we get the model:

$$\mathrm{P}(Y_{i,0:t} \mid X_{i,t'\in\mathbb{Z}}) = \mathrm{P}(Y_{i,0} = y_{i,0} \mid X_{i,t'\in\mathbb{Z}}) \prod_{s=1}^{t} \mathrm{P}(Y_{i,s} = y_{i,s} \mid Y_{i,0:s-1}, X_{i,t'\in\mathbb{Z}})$$

It's easy to show that under the assumptions above the conditional probabilities on the right hand side of the equation can be reduced to depend only on the event just before due to the Markov property of time-to-event models. It is a model assumption that $P(DBB = T \mid X) = P(Y \mid X)$ so that the probability of DBB depends on the covariates (Song, 2010, ch 2.3.1). We get the simplified model:

$$\mathrm{P}(Y_{i,0:t} \mid X_{i,t'\in\mathbb{Z}}) = \mathrm{P}(Y_{i,0} = y_{i,0} \mid X_{i,t'\in\mathbb{Z}}) \prod_{s=1}^{t} \mathrm{P}(Y_{i,s} = y_{i,s} \mid Y_{i,s-1} = y_{i,s-1}, X_{i,t'\in\mathbb{Z}})$$

The conditional probability of bud burst at time t_i is:

$$\begin{aligned}
\mathrm{P}(T_i = t_i \mid X_i) &= \mathrm{P}(Y_{i,0} = Y_{i,1} = \ldots = Y_{i,t_i-1} = 0, Y_{i,t_i} = Y_{i,t_i+1} = \ldots = 1 \mid X_i) \\
&= \{\text{since } \mathrm{P}(Y_{i,t_i+\tau} = 1 \mid Y_{i,t_i+\tau-1} = 1) = 1 \; \forall \tau \geq 0\} \\
&= \mathrm{P}(Y_{i,0} = Y_{i,1} = \ldots = Y_{i,t_i-1} = 0, Y_{i,t_i} = 1 \mid X_i) \\
&= \{\text{using the markov property and inital assuption of } P(Y_{i,0} = 0) = 1\} \\
&= \left[\prod_{s=1}^{t_i-1} \mathrm{P}(Y_{i,s} = 0 \mid Y_{i,s-1} = 0, X_i) \right] \cdot \mathrm{P}(Y_{i,t_i} = 1 \mid Y_{i,t_i-1} = 0, X_i)
\end{aligned}$$

It is clear in this case that the covariates can only be influencing up to their present time so that $X_{i,0:s}$ should replace X_i in the above derivation, and $\mathrm{P}(Y_{i,0}=0 \mid X_i)=1$ by assumption (before January 1 there is no budburst).

Now assuming the transition probabilities can be written as equation:

$$P_t = P(Y_t=1|Y_{t-1}=0,\mathrm{X}) = \mathrm{logit}^{-1}(X\beta) = \frac{1}{1+e^{-X\beta}}$$

then the probability of having budburst at time $T=t$ is:

$$P_t = P(Y_t = 1 \mid Y_{t-1} = 0)$$
$$P(T=1) = P_1$$
$$P(T=t) = \prod_{s=1}^{t-1}(1-P_s)\cdot P_t \quad , \; t>1$$

And the probability of having budburst not later than $T=t$ is:

$$P(T \le t) = \sum_{u=1}^{t} P(T=u) = P_1 + \sum_{u=2}^{t}\prod_{s=1}^{u-1}(1-P_s)\cdot P_u$$

3.2 Regression

The method of linear regression is commonly used to model the dependence between explanatory variables and a response variable under analysis (Rawlings, 2001). Least squares (LS) minimization is applied to the coefficients of the explanatory variables in order to minimize the squared error of the fitted model and the original data. The residual error of the model is assumed to be standard normal Gaussian with mean 0 and variance σ^2. The setup of multiple linear regression is:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_p X_{pi} + e_i$$

The explaining variables can have any form: continuous, constant (the intercept being the default constant equal to one) or categorical. However the response Y must be continuous in order for the least squares to make sense. So in the case of a categorical response having values of zeros and ones standing for the states true and false linear regression cannot be applied.

3.3 Multiple logistic regression

In the case of a binary response $Y_i \in \{0,1\}$ as for the Markov chain (see figure 3) described in 3.1 above, Y_i is distributed as $Bin(1,n) = Bernoulli(n)$. Here (0, 1) stands for (DBB not occurred, DBB occurred). It is interesting to model the probability p_i of success given explanatory variables X. Trying

$$p_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_{pp} X_{ip} = X_{i\cdot}\beta$$

fails since $p_i \in [0,1]$ and we cannot guarantee that $p_i = X_{i\cdot}\beta \in [0,1]$. The transformation

$$\mathrm{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right)$$

maps the probabilities to the real line and linear regression can be applied. This mapping function is called the logit-*link function*. However Y is not continuous and therefore the method of maximum likelihood (ML) is used to find optimal coefficients. The logit transformation is natural since it defines the natural parameter η for the corresponding *exponential family* (Andersen, 1970) which the binomial distribution belongs to. Thus there exist *sufficient statistics* (Casella, 2002) and there are *conjugate priors* (Gelman, 2003) in Bayesian analysis. The model will for each observation fit a value on the real line and the transformation, using the inverse *logistic function*:

$$P(x) = \text{logit}^{-1}(x) = \frac{1}{1+e^{-x}}$$

converts the values back to probabilities. So the model:

$$\text{logit}(p_i) = \beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip} = X_{i\cdot}\beta$$

inverted with the logistic function gives

$$p_i = \frac{1}{1+e^{-X_{i\cdot}\beta}} = \frac{e^{X_{i\cdot}\beta}}{e^{X_{i\cdot}\beta}+1}$$

which can be estimated by maximizing the binomial likelihood (Christensen, 1990, ch 2.6):

$$L = \prod_{i=1}^{N} \binom{1}{y_i} p_i^{y_i}(1-p_i)^{1-y_i} = \prod_{i=1}^{N} \left(\frac{e^{X_{i\cdot}\beta}}{e^{X_{i\cdot}\beta}+1}\right)^{y_i} \left(\frac{1}{e^{X_{i\cdot}\beta}+1}\right)^{1-y_i} = \prod_{i=1}^{N} \frac{\left(e^{X_{i\cdot}\beta}\right)^{y_i}}{e^{X_{i\cdot}\beta}+1}$$

Here $\boldsymbol{y_i}$ is the response for observation vector $X_{i\cdot}$. Another change compared to linear regression is that the residuals are non-Gaussian due to the transformation. The maximization of the likelihood estimates β and we get the transition probabilities using the logistic function on the prediction covariates multiplied with β as described at the end of section 3.1. See also (Song, 2010, ch 2.3.3)

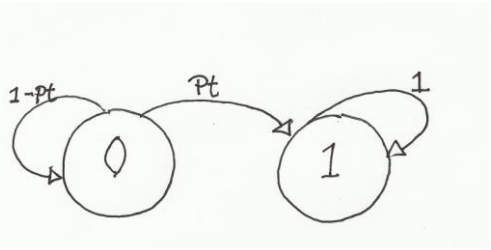

Figure 3 Two state Markov transition model. The Markov chain is in the 'zero' state and remains there at each timepoint t with probability $1-P_t$. Consequently the chain changes state with probability P_t at each timepoint t. When the markov chain has changed state to the 'one' state it will remain there with probability 1 and has thus been absorbed by the 'one' state.

4. Variable selection by regularization

When fitting a regression model with many covariates linear regression or logistic regression may over fit the model. To over fit means that the model will readily explain the data used to estimate parameters but will do poorly trying to predict responses given new covariate data. This happens since more explanatory variables will increases the fit, and thereby the likelihood. Common methods to resolve the over fitting problem is to penalize the likelihood of the model for having many parameters allowing a compromise between model complexity and fit. This gives rise to information criteria upon which to choose between different models. One usually uses the Akaike information criteria (AIC) or the Bayesian information criterion (BIC). See (Brokwell, 1991).

Another way of handling over fitting is to use *cross validation* (CV) (Picard, 1984). Part of the data used to estimate parameters are used to estimate the prediction error of the model. This can then be used to differentiate between models. Additional validation data is then used to assess the accuracy of predictions. These procedures can be applied to cases where a few models are being compared. Otherwise one will have to manually compare many different models where the estimated models can change drastically if some covariate(s) are added/removed, making it hard to pinpoint the most relevant factors. Effectively one will have to check all combinations often making the methods computationally intractable. There are also situations where covariates cancel each other out. They can at the same time be good predictors in the model and have the same amount of relevance.

Dealing with hundreds or more covariates the problem of variable selection can be (more) objectively solved with regularization of the regression. By adding a penalty in the regression model to the coefficient sizes the optimization will shrink some coefficients quicker to zero and by applying CV the penalty term which maximizes the out of sample prediction error can be determined. This maximization at the same time reduces the number of explaining variables by minimizing the coefficients (one typically chooses a cutoff point for the regression coefficients when to discard a covariate or as in the case of lasso, coefficients automatically shrink to zero). Hopefully the covariates that remain in the model will be few enough to make good predictions and many enough to, at the same time, explain existing data well.

4.1 Ridge regression

One common regularization method is ridge regression, or Tikhonov regularization (Dykes, 2012). This adds the sum of the squared coefficients as a penalty. Correlated coefficients with high variance are thereby prevented from cancelling each other by having coefficients with different signs.

$$\hat{\beta}^{ridge} = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$$

As seen the intercept is not penalized to avoid dependence on the origin of Y. Looking closer on this expression the first term is the negative log likelihood of a multivariate Gaussian distribution which is the setup for linear regression having independent observations. The second term is the negative log likelihood of a Bayesian prior distribution.

$$p(\beta) = \exp\left(-\lambda \sum_{j=1}^{p} \beta_j^2 \right)$$

Regarding the coefficients as independent, this introduces a prior distribution precisely with $\beta_j \sim N(0,1/(2\lambda))$.

4.2 Lasso

The lasso (Tibshirani, 1996) is slightly different from ridge in that it penalizes the sum of the magnitudes of the coefficients instead of their squares. It has the advantage of forcing small coefficients to zero, eliminating the need of choosing a cutoff threshold.

$$\hat{\beta}^{lasso} = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \qquad (1)$$

The lasso has no closed form solution but it is a quadratic programming problem and related to convex optimization problem (Hastie, 2009, ch 3.4.2) and efficient algorithms exists. Lasso also works for logistic regression since the regularization can be viewed as setting a prior on the coefficients. For lasso the prior is $p(\beta) = e^{-\lambda|\beta|}$, or a laplace distribution. See (Tibshirani, 1996, ch 8)

Cross validation

The penalty term λ in (1) needs to be specified and it's not obvious which value to choose. The solution is to apply cross validation to the data set. One divides the set into k sub partitions and performs the estimation on all but one set. The remaining set is used for validation. This procedure is repeated for each partition. The λ parameter is then selected as what minimizes the deviance or root mean square error (RMSE), to maximize the likelihood of the validation set. At minimum $\lambda_{\min}$ those variables not set to zero are the datasets best predictors. This set is then further reducible by increasing the lambda parameter so much so that the error of the minimum still overlapped by one standard deviation λ_{1se}. The two models ($\lambda_{\min}$ and λ_{1se}) are still within statistical significance as the error estimates of the cross validation overlap to 50% (=1se). So from the point of view of the cross validation the two λ values are statistically overlapping (or there's a 50% chance they are equal).

5. Analysis

5.1 R – Package

The open source statistical computing environment R has been used for this analysis. The R-package *glmnet* (Friedman, 2010) provides tools for lasso analysis of among other logistic-regression. The covariates for each station are pre-computed and saved for efficiency. The R-package *maps* (Becker, 2012) is used when illustrating the data. Tables printed using the *gplots* (Warnes, 2012) package.

5.2 Covariates

The main aim of the analysis is to find relevant covariates determining the DBB for birch in Finland and the United Kingdom. The covariates are divided into two groups, constant and varying. The constant covariates are fixed for each year and station such as: location or mean temperature the previous year. The varying covariates are temperature, GDD etc. Along with these covariates all interactions between the time constant and varying are considered. An interaction between two covariates is simply the multiplied value of the two covariates. In this case it will capture dependence on the previous year and/or location. Some conditions might inhibit budburst and others promote. See the appendix for a complete list of covariates. In this case we have with interactions $11 \cdot 17 + 11 + 17 = 215$ covariates to choose from.

5.3 Data selection

For the regression to work all covariates must exist. A much more complicated model could of course handle the case of missing covariate values, but given the abundance of data the added complication is not justifiable. So the available stations for a full analysis are considered. In general 2/3 of the locations are used for analysis and 1/3 for validation. For the United Kingdom the number of DBB stations is far more than the number of climate cells. Therefore the climate cells are used and a smaller subset is randomly selected. The DBB stations associated with the climate cells are used in analysis. It turns out that too much data only over fits the model. The three models (UK3 a subset of UK2 beeing a subset of UK1) analyzed are:

UK1 – 33% of the climate cells associated.
UK2 – 17% of the climate cells associated.
UK3 – 11% of the climate cells associated.

5.4 Parameter estimation

For the cross validation to work properly the data needs to be grouped according to the observations which in this case are a set of days (1 January up to the DBB) instead of the default random grouping of Y. This is done by supplying an nfoldid parameter to glmnet, which indexes, for each observation, what CV-group it should belong to, figure 5.4a illustrates with an example.

```
> regression.data$foldid[1:900]
  [1] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [60] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[119] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 9
[178] 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9
[237] 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9
[296] 9 9 9 9 9 9 9 9 9 9 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7
[355] 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7
[414] 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 4 4 4 4 4 4 4 4 4 4 4 4
[473] 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
[532] 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
[591] 4 4 4 4 4 4 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8
[650] 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8
[709] 8 8 8 8 8 8 8 8 8 8 8 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5
[768] 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5
[827] 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[886] 3 3 3 3 3 3 3 3 3 3 3 3 3 3
```

Figure 5.4a nFoldid structure marking the individual observations for 10 folds crossvalidation. Note that the 10 groups are randomly selected, but the 0/1 sequence associated with one DBB observation is not split between groups.

The parameters estimated through the lasso regularization are those constrained by penalization. Thus estimating them using only GLM should give slightly larger values. In addition parameter standard errors are also obtained.

Finland

The stations for analysis are shown in figure 2a and after lasso cross validation the minima and one standard error reduction are shown in figure 5.4b. The minima minimize the cross validation error but do so having a very large set of covariates in the model. Intuitively if a point on the curve has it's CI overlapping half that of the minima it is statistically close. This is where the λ_{1se} point lies, and one chooses λ_{1se} to ensure the largest possible reduction in the number of covariates. Figure 5.4c shows the coefficient trajectories as a function of λ, note the shrinkage to 0 for large penalties.

Table 5.4d shows the covariates for the λ_{1se} model and coefficients before and after re-estimation. Clearly three covariates doesn't seem to have significance in the model, but it is not clear how to reduce the model further. By insignificant means that the standard error is large compared to the coefficient estimate. In this case the standard error is as big, or almost, as the coefficient.

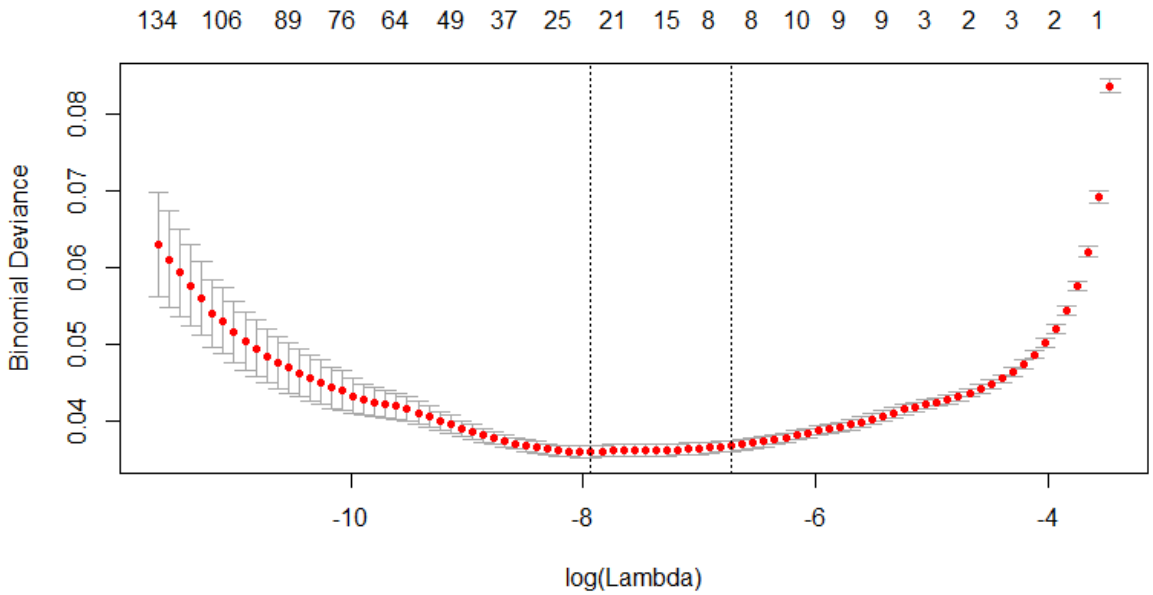

Figure 5.4b Cross validation for binomial logistic regression with lasso regularization on Finland data. At minimum there are 22 relevant coefficients (including the intercept) and one standard error away that number is reduced to 8. The reduction is not monotone.

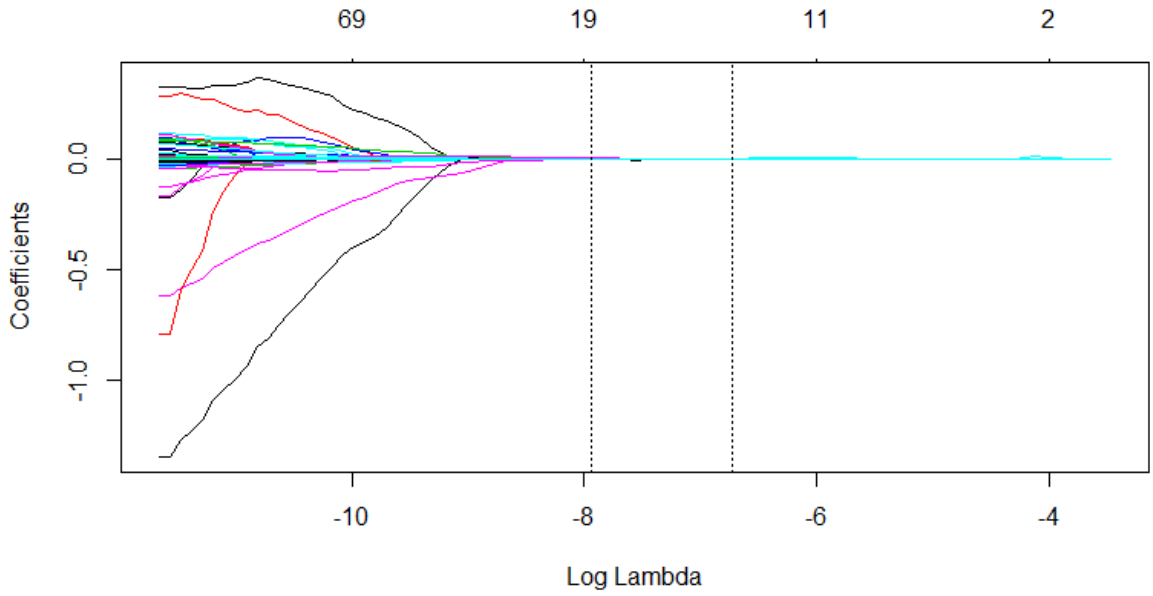

Figure 5.4c The model coefficients for Finland are forced to zero as λ gets larger (varying from just above zero to one). The lines for the $\lambda_{\min}$ and λ_{1se} are as in figure 5.4b

Table 5.4d After 10 fold cross validation these are the one standard deviation λ_{1se} estimated parameters for the Finland data and shown are the estimated parameters using GLM only. As seen all the coefficients are larger in magnitude. This is due to the lack of the penalizing term. The covariates 4, 5 and 7 have p-value less than 90% significance.
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

FI (.1se) covariates after 10 fold CV and refit with GLM

	covariate	coeff (cv)	coeff (glm)	std.err (glm)	Pr(>|z|)	sign.
[1,]	(Intercept)	-9.06e+00	-1.26e+01	5.24e-01	7.87e-128	***
[2,]	total.fall	-1.00e-03	-5.19e-03	1.82e-03	4.35e-03	**
[3,]	continentality.cig:agdd0	6.75e-04	1.34e-03	8.96e-05	7.20e-51	***
[4,]	latitude:agdd5	3.17e-04	3.87e-04	3.02e-04	1.99e-01	
[5,]	continentality.cig:agdd5	2.72e-04	4.57e-04	4.88e-04	3.49e-01	
[6,]	chill5:agdd5	2.46e-04	3.57e-04	1.57e-04	2.32e-02	*
[7,]	mean.year:lastfrost	2.57e-03	3.07e-03	2.25e-03	1.71e-01	
[8,]	height.dbb:growthseason	-1.11e-04	-2.11e-04	2.91e-05	4.80e-13	***

The United Kingdom

The stations for analysis of the UK3 model are shown in figure 2c. After lasso cross validation the minimum $\lambda_{\min}$ and one standard error reduction of the model λ_{1se} are shown in figure 5.4i. The coefficient trajectories are shown in figure 5.4j. Table 5.4e shows the estimated coefficients of the lasso regularization along with the glm re-estimated coefficients for the UK3 λ_{1se} model. As seen the covariate latitude:agdd5 change sign at re-estimation although all coefficients have significance. For comparison the covariates for the UK2 model are shown in table 5.4f. There are five covariates in the UK2 model which are missing in the UK3 model. In addition we see that the covariate latitude:agdd5 is new to the UK3 model.

The covariate latitude:agdd5 is therefore removed from the UK3 model and the re-estimated coefficients are as in table 5.4g. This new model, called (UK3 re-estimated reduced), has the covariate chill5:agdd0 with no significance (large standard error). Removing the covariate chill5:agdd0 and creating the model, UK3 re-estimated reduced 2, all parameters have significance (see table 5.4h). These two reduced models again gives good predictions only slightly worse than for the other models. The model in table 5.4h is the smallest and most efficient containing only 5 covariates.

Table 5.4e After 10 fold cross validation these are the one standard deviation (.1se) estimated parameters for the UK3 data and shown are the estimated parameters using GLM only. As seen most of the coefficients are larger in magnitude. One have however changed sign, marked in blue. It could mean the model is further reducible. Note that the model gives standard errors for the coefficient estimations.
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

UK (.1se) covariates after 10 fold CV and refit with GLM

	covariate	coeff (cv)	coeff (glm)	std.err (glm)	Pr(>|z|)	sign.
[1,]	(Intercept)	-1.51e+01	-2.75e+01	2.82e-01	0.00e+00	***
[2,]	daylength	7.31e-01	1.70e+00	2.15e-02	0.00e+00	***
[3,]	agdd5	1.21e-03	5.11e-02	4.91e-03	2.54e-25	***
[4,]	chill5:agdd0	9.34e-06	2.36e-05	3.53e-06	2.39e-11	***
[5,]	latitude:agdd5	4.36e-08	-8.11e-04	8.87e-05	6.45e-20	***
[6,]	continentality.cic:agdd5	5.11e-04	4.20e-04	6.12e-05	7.10e-12	***
[7,]	total.spring:agdd5	1.45e-05	2.54e-05	4.31e-06	3.60e-09	***
[8,]	number.spring:agdd5	9.94e-05	1.29e-04	2.11e-05	1.14e-09	***

Table 5.4f The UK2 model after cross validation. There are six covariates, colored red, that isn't in the UK3 model (see the preceeding table 5.4e). In addition the covariate for the UK3 model that changes sign at re-estimation is new compared to the UK2 model.

UK (.1se) covariates after 10 fold CV and refit with GLM

	covariate	coeff (cv)	coeff (glm)	std.err (glm)	Pr(>|z|)	sign.
[1,]	(Intercept)	-1.66e+01	-2.80e+01	3.12e-01	0.00e+00	***
[2,]	daylength	8.23e-01	1.74e+00	3.47e-02	0.00e+00	***
[3,]	agdd0	2.86e-04	-5.50e-03	1.32e-03	3.15e-05	***
[4,]	agdd5	7.35e-05	9.59e-03	4.05e-03	1.79e-02	*
[5,]	continentality.cic:temperature	1.45e-04	1.95e-03	1.48e-03	1.85e-01	
[6,]	chill5:temperature	3.02e-06	-5.82e-04	4.04e-04	1.50e-01	
[7,]	continentality.cic:agdd0	7.60e-05	4.79e-04	8.56e-05	2.20e-08	***
[8,]	chill5:agdd0	1.13e-05	4.34e-05	7.78e-06	2.52e-08	***
[9,]	continentality.cic:agdd5	3.93e-04	-3.97e-04	2.56e-04	1.21e-01	
[10,]	mean.year:agdd5	7.45e-05	-1.37e-03	2.47e-04	3.07e-08	***
[11,]	mean.fall:agdd5	2.15e-06	2.43e-03	2.51e-04	3.25e-22	***
[12,]	total.spring:agdd5	1.89e-05	1.96e-05	3.82e-06	2.91e-07	***
[13,]	number.spring:agdd5	2.59e-05	1.69e-04	2.22e-05	2.47e-14	***

Table 5.4g The UK3 re-estimated and reduced model. No coefficients change sign. Predictions are good. The covariate chill5:agdd0 can be dropped and the resulting model have all coefficients significant see table 5.4h.

UK (.1se) covariates after 10 fold CV and refit with GLM

	covariate	coeff (cv)	coeff (glm)	std.err (glm)	Pr(>|z|)	sign.
[1,]	(Intercept)	-1.51e+01	-2.75e+01	2.96e-01	0.00e+00	***
[2,]	daylength	7.31e-01	1.72e+00	2.34e-02	0.00e+00	***
[3,]	agdd5	1.21e-03	7.62e-03	8.40e-04	1.12e-19	***
[4,]	chill5:agdd0	9.34e-06	4.13e-06	2.90e-06	1.55e-01	
[5,]	continentality.cic:agdd5	5.11e-04	7.13e-04	5.70e-05	6.65e-36	***
[6,]	total.spring:agdd5	1.45e-05	2.50e-05	4.40e-06	1.34e-08	***
[7,]	number.spring:agdd5	9.94e-05	9.15e-05	2.12e-05	1.51e-05	***

Table 5.4h The further reduced UK3 model. All coefficients are significant. This is the most efficient model for the UK data set.

UK (.1se) covariates after 10 fold CV and refit with GLM

	covariate	coeff (cv)	coeff (glm)	std.err (glm)	Pr(>|z|)	sign.
[1,]	(Intercept)	-1.51e+01	-2.76e+01	2.83e-01	0.00e+00	***
[2,]	daylength	7.31e-01	1.73e+00	2.09e-02	0.00e+00	***
[3,]	agdd5	1.21e-03	7.91e-03	8.10e-04	1.73e-22	***
[4,]	continentality.cic:agdd5	5.11e-04	7.09e-04	5.64e-05	3.62e-36	***
[5,]	total.spring:agdd5	1.45e-05	2.34e-05	4.24e-06	3.42e-08	***
[6,]	number.spring:agdd5	9.94e-05	9.58e-05	2.07e-05	3.76e-06	***

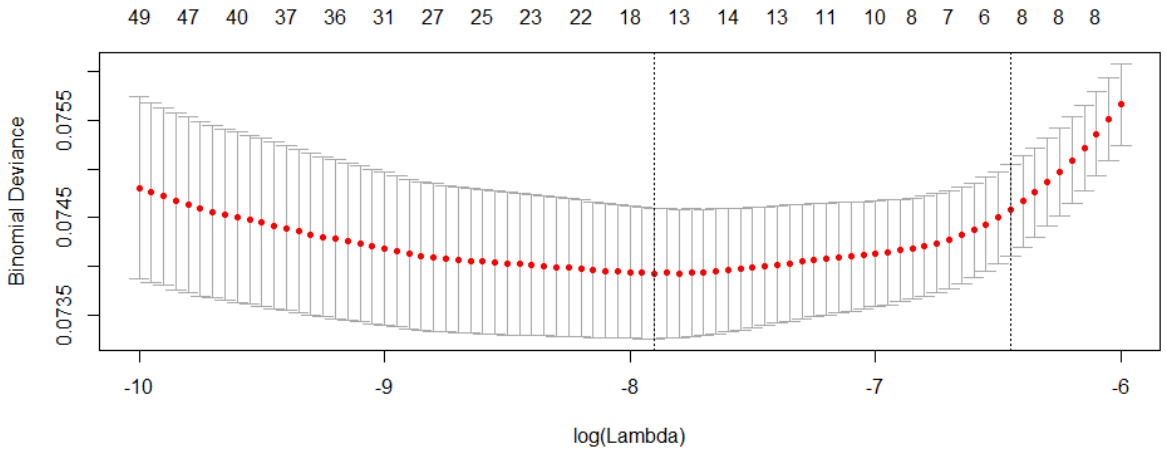

Figure 5.4i Cross validation for binomial logistic regression with lasso regularization on UK3 data. At minimum there are 17 relevant coefficients (including the intercept) and one standard error away that number is reduced to 8. The reduction is not monotone.

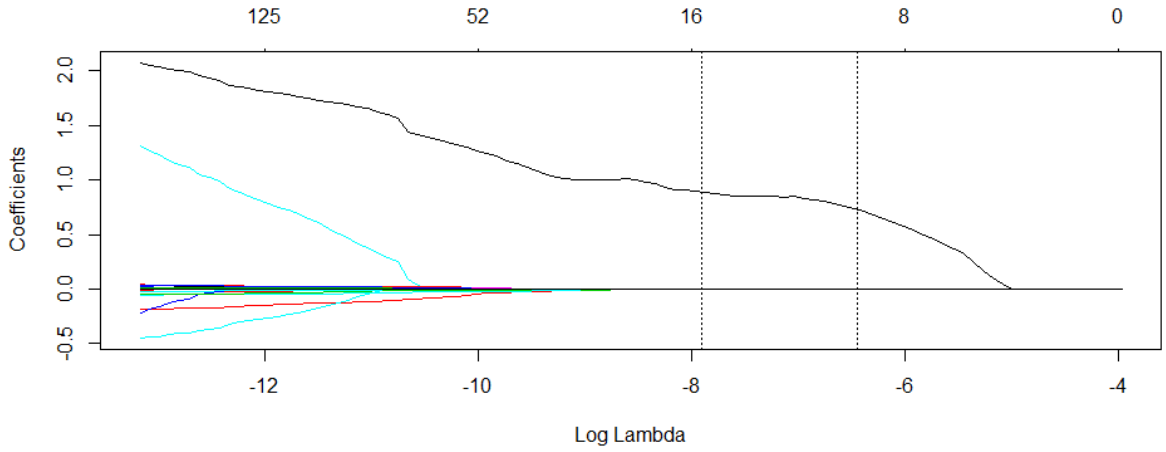

Figure 5.4j The coefficients are forced to zero as λ gets larger (varying from just above zero to one). The lines for the $\lambda_{\min}$ and λ_{1se} are as in figure 5.4i.

The above curve is for day length. It might model a linear accumulated state in the response. One day one unit. For the Finland data with no local variation it is not represented.

5.5 Predictions

The predictions are made with both the lasso coefficients and the re-estimated coefficients using generalized linear model (GLM) for the validation data. Two types of prediction intervals (PI) are formed:

1. The accumulated probability PI is described in chapter 3.3 and is the probability that the budburst occurs at a specific day given the previous days of the year. See figure 5.5h for an example.

2. The transition probability PI is the regression probability for each day of having budburst individually. See figure 5.5i for an example.

The PI is formed as centered at the day closest to having 50% probability (for respective interpretation) and the upper 97.5% and lower 2.5% respectively. It turns out that the re-estimated model optimizes predictions at the 50% point of the second PI and that the lasso regularized model at the 50% point of the first PI.

The two data sets are very different as the DBB information for Finland has no local variation (only one DBB value per climate cell and year). Se figure 2e-g. As a consequence the prediction get more accurate for Finland. Overall the prediction results are good as seen in table 5.5f. Comparing with (Song, 2010, table 4.4) both the MAE and RMSE is significantly reduced. The results are also comparable to the sequential-i model evaluated in (Kramer, 1994, table 2). It is clear that the $\lambda_{\min}$ min model is comparable to the λ_{1se} model. The simpler λ_{1se} model can therefore be used without loss of predictive accuracy.

The predictions together with the observed DBB are showed graphically in figure 5.5a-d. The UK model doesn't take into account DBB observations below around 85 and above 125 as seen in figure 5.5b. They are treated as extreme observations and are rare. So rare that they might be errors in the data set but that is out of the scope of this analysis. Clearly it is important that the modeled data is correct and doesn't vary too much if the predictions are to be accurate.

For reference the predictions on the regression data set is in table 5.5g. The values are only very slightly better than for the validation set. It is also clear the for the UK data too much stations in the analysis creates over fit and the model becomes overcomplicated without gaining in prediction accuracy. The model UK3 with about 11% of the stations gives 8 covariates including the intercept and gives as good prediction result as the other models. The model can be further reduced to only 5 covariates and the intercept.

The prediction interval both for the Finland data and the UK data are too wide. The problem is in finding a better way of forming the PI in a consistent way. The extreme values of the UK

data, in this case less than 80 and more than 130 in DBB, make up about 2.7% of the validation data. That could account for the slightly low value of the PI cover of around 93% for the re-estimated UK models. For Finland it is clear that the lower part of the PI is far too low. It is though not trivial to adjust the lower limit of the PI. One 'quick-fix' would be to measure the length between the upper limit and the prediction point and then subtract a scaled amount from the prediction to get the lower limit. Doing this for a scaling of 2 actually brings the average PI length down to about 14 days for the Finland data see figure 5.5e. Doing the same but with a scaling of 1 on the UK3 data reduces the average PI length from 64 days to 37.2 days and a PI cover of 93.6% comparable to the UK re-estimated models.

Trying to put an upper limit for the PI will take down the PI cover by the extreme values. Imagine putting a horizontal line somewhere in the red band in figure 5.5b around 130 and 80, which is the interval for all predictions. It is clear that then the extreme values will pull down the PI cover. But since the extreme values aren't modeled, or at least seem ignored by the predictions, it is not really possible to anticipate them, and therefore the PI length cannot be made smaller. The model makes a weighted choice and therefore the predictions are close to the mean values and hence the extreme values

For the UK data additional errors are calculated as:

(1) $\sum_i |T^i_{obs} - T^i_{median,obs}|$, the theoretical best MAE error.

(2) $\sum_i (T^i_{obs} - T^i_{mean,obs})^2$, the theoretical best RMSE error.

(3) $\sum_i |T^i_{pred} - T^i_{mean,obs}|$, the predicted DBB against observed mean for MAE estimate.

(4) $\sum_i (T^i_{pred} - T^i_{mean,obs})^2$, the predicted DBB against observed mean for RMSE estimate.

The error estimates (1) and (2), calculate the optimal MAE and RMSE for data having multiple values that is to be estimated with one value. For the MAE being in the L1-norm this one value that minimizes the error is the median of the values to estimate (or in this case the median of the DBB for a climate cell and year pair). For the RMSE the value that minimizes the climate cell/year observations is the mean of the observations.

The error estimates (3) and (4) calculate the errors between the predictions and the mean of the climate cell/year observations. That would be the errors if one would form the averages of the observations and thereby removing local variation. Technically there is a difference in taking the averages before analysis and then doing predictions.

The error estimates (1) and (2) are zero for the Finland data since observed DBB coincides with median and mean. The cases (3) and (4) coincide with the ordinary MAE and RMSE estimates for Finland.

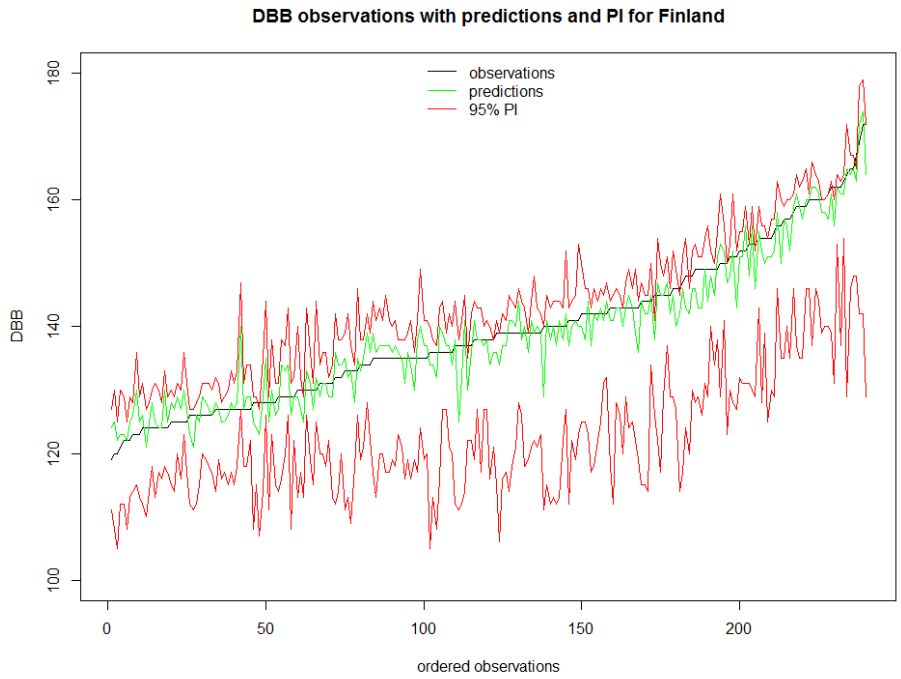

Figure 5.5a The predictions compared to the observations for the Finland data together with prediction upper and lower intervals. Note that the low part is shifted downward.

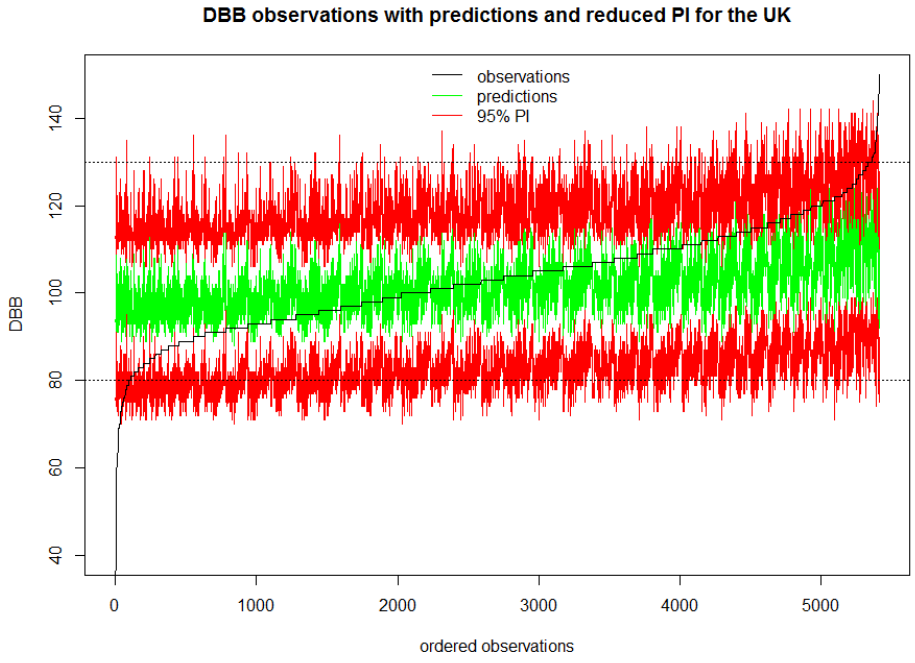

Figure 5.5b The predictions compared to the observations for the UK data together with prediction upper and lower intervals. Due to large local variation the predictions with PI are irregular. Note that values of DBB below around 80 and above 130 are not modeled at all.

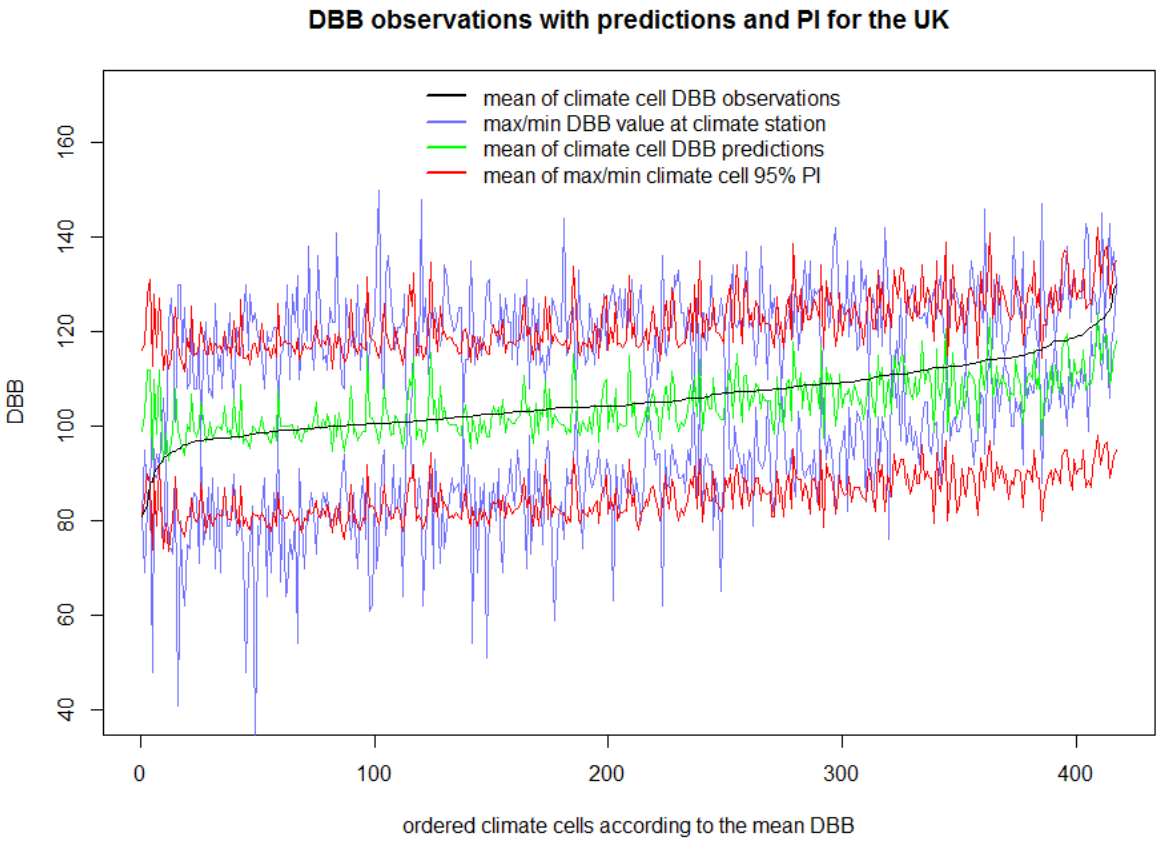

Figure 5.5c Predictions compared to observation according to the climate cells. The gray lines are the maximum respectively the minimum observations at each climate cell. The curves vary very much since even the variation across years are taken into account (see Figure 5.5d for the year 2001 specifically). The red PI curves are the average max/min PI for each climate cell.

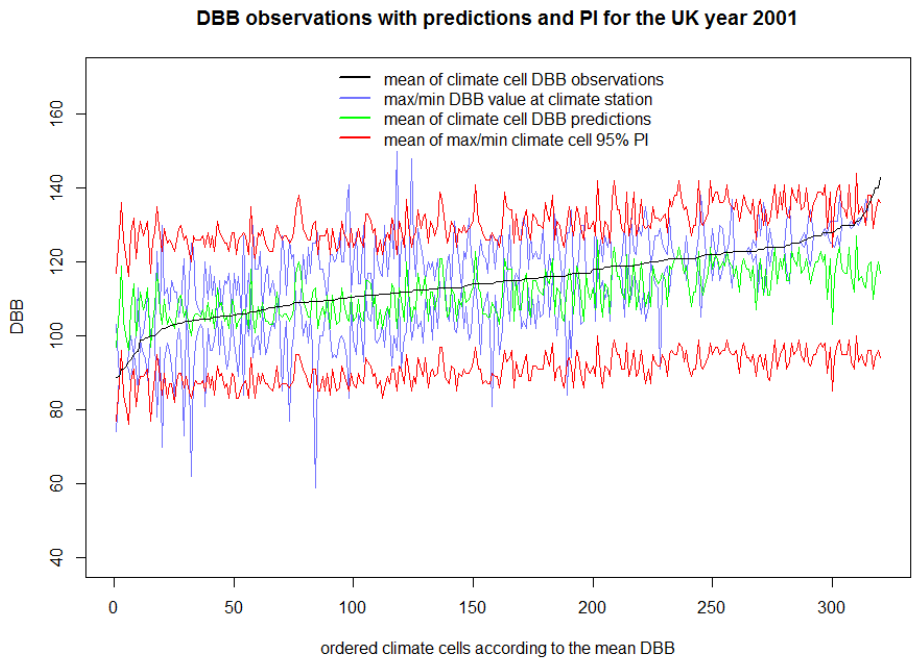

Figure 5.5d Predictions compared to observation according to the climate cell means for the UK data year 2001.

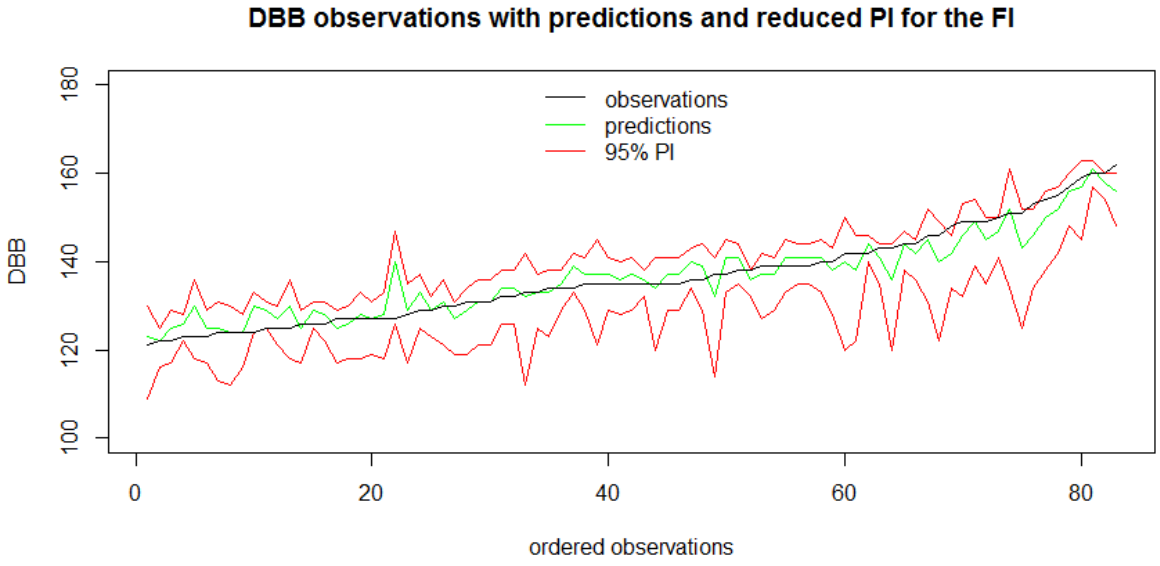

Figure 5.5e The predictions (on the validation set) together with observations and upper and lower PI curves. The lower PI part is formed as the distance between the predictions (green curve) and the upper PI part times 2 taken below the prediction curve. The average length of the PI cover reduces from 20.9days to 14.1days without changing the PI cover.

Table 5.5f Prediction results on the ***validation data*** before and after coefficient re-estimation. The MAE and RMSE for the error estimates (1)-(4) are included. For (3) and (4) the errors are actually smaller since the logistic regression tends to approximate the original DBB data closer to the cell average. The MAE for the optimal prediction (1) is taken with respect to the median not the average since it is the minimizer in the L1 norm. The best models are marked in blue. The UK models have 33%, 17% and 11% of the available gridded climate cells for model estimation. Clearly too many stations doesn't add to prediction and only over fit the model. The UK3 re-estimated and reduced models is the most compact and efficient (among the analyzed) model for the UK data.

	MAE	RMSE	Bias	95 % PI coverage	95 % PI length	Model size	MAE (1)	RMSE (2)	MAE (3)	RMSE (4)
Finland .1se	2.51	3.33	0.1	92.8%	20.9	8				
Finland .min	2.59	3.56	0.3	84.3%	13.9	22				
Finland re-estimated	2.42	3.34	0.2	85.5%	11.0	8				
UK1 .1se	7.74	10.06	0.2	95.2%	52.2	17	5.42	7.96	4.44	6.16
UK1 .min	7.53	9.85	0.2	94.7%	52.2	42	5.42	7.96	4.29	5.81
UK1 re-estimated	7.68	10.02	-0.3	93.9%	37.6	17	5.42	7.96	4.38	6.09
UK2 .1se	7.75	10.11	-0.4	95.5%	57.0	13	5.40	7.96	4.54	6.23
UK2 .min	7.76	10.10	0.3	94.6%	49.5	13	5.40	7.96	4.49	6.21
UK2 re-estimated	7.71	10.09	-0.4	93.2%	37.1	13	5.40	7.96	4.51	6.19
UK3 .1se	7.77	10.13	-0.8	95.9%	64.0	8	5.38	7.95	4.60	6.28
UK3 .min	7.76	10.09	0.3	94.7%	50.4	17	5.38	7.95	4.50	6.21
UK3 re-estimated	7.70	10.05	-0.3	93.3%	37.0	8	5.38	7.95	4.51	6.15
UK3 re-est.& red. 1	7.71	10.07	-0.3	93.6%	36.8	7	5.38	7.95	4.55	6.19
UK3 re-est.& red. 2	7.82	10.21	-1.7	92.6%	36.4	6	5.38	7.95	4.76	6.41

Table 5.5g Prediction results on the ***regression data*** (data that estimated the coefficients) before and after coefficient re-estimation. The values are only slightly better than those for the validation set.

	MAE	RMSE	Bias	95 % PI coverage	95 % PI length	Model size	MAE (1)	RMSE (2)	MAE (3)	RMSE (4)
Finland .1se	2.42	3.18	-0.6	90.4%	21.6	8				
Finland .min	2.27	2.95	-0.4	86.7%	14.0	22				
Finland re-estimated	2.50	3.33	-0.3	88.5%	10.9	8				
UK1 .1se	7.53	9.86	0.2	94.7%	52.2	17	5.42	7.96	4.29	5.81
UK1 .min	7.49	9.78	0.7	94.1%	45.4	42	5.42	7.96	4.17	5.70
UK1 re-estimated	7.47	9.80	-0.2	93.9%	37.5	17	5.42	7.96	4.18	5.72
UK2 .1se	7.41	9.69	-0.3	95.6%	57.1	13	5.40	7.96	4.10	5.59
UK2 .min	7.36	9.59	0.3	94.8%	72.3	13	5.40	7.96	3.98	5.42
UK2 re-estimated	7.33	9.65	-0.2	94.1%	37.1	13	5.40	7.96	4.04	5.52
UK3 .1se	7.40	9.70	-0.8	96.4%	64.2	8	5.38	7.95	4.07	5.50
UK3 .min	7.33	9.59	0.3	94.3%	50.5	17	5.38	7.95	3.87	5.30
UK3 re-estimated	7.31	9.64	-0.2	93.8%	37.2	8	5.38	7.95	3.94	5.39
UK3 re-est.& red. 1	7.33	9,68	-0.2	93.4%	36.8	7	5.38	7.95	3.93	5.46
UK3 re-est.& red. 2	7.45	9.83	-1.5	92.5%	36.4	6	5.38	7.95	4.25	5.72

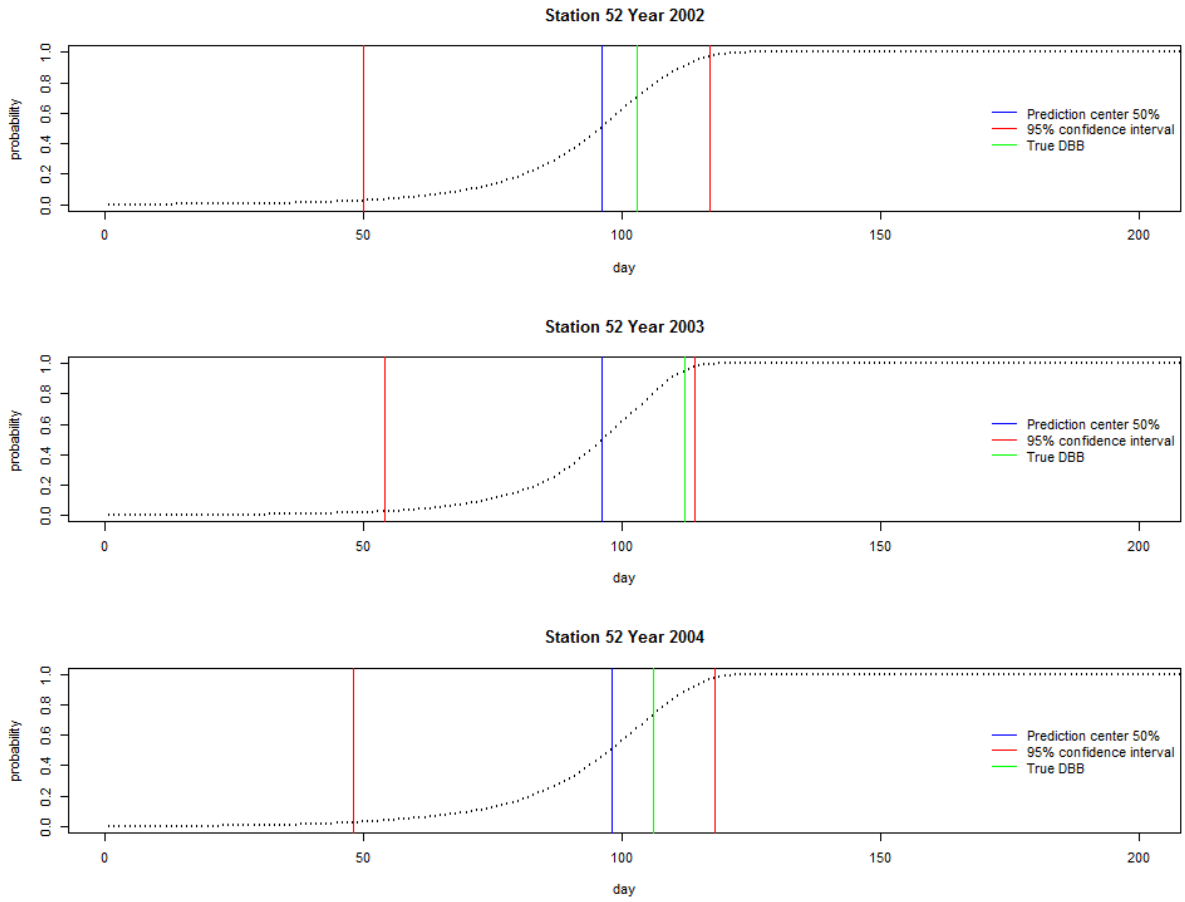

Figure 5.5h Predictions on the UK3 .1se model with the 50% probability marked with blue line. That is the point where it is as much probable for a DBB to have occurred as it is to not have occurred. This is the first variant of the PI described in 5.4. The 95% prediction cover is marked with red lines. The true DBB days are marked with green lines. Note the stretching of the cumulative distribution function (CDF) curve (the dotted curve) making the left part of the PI land farther to the left.

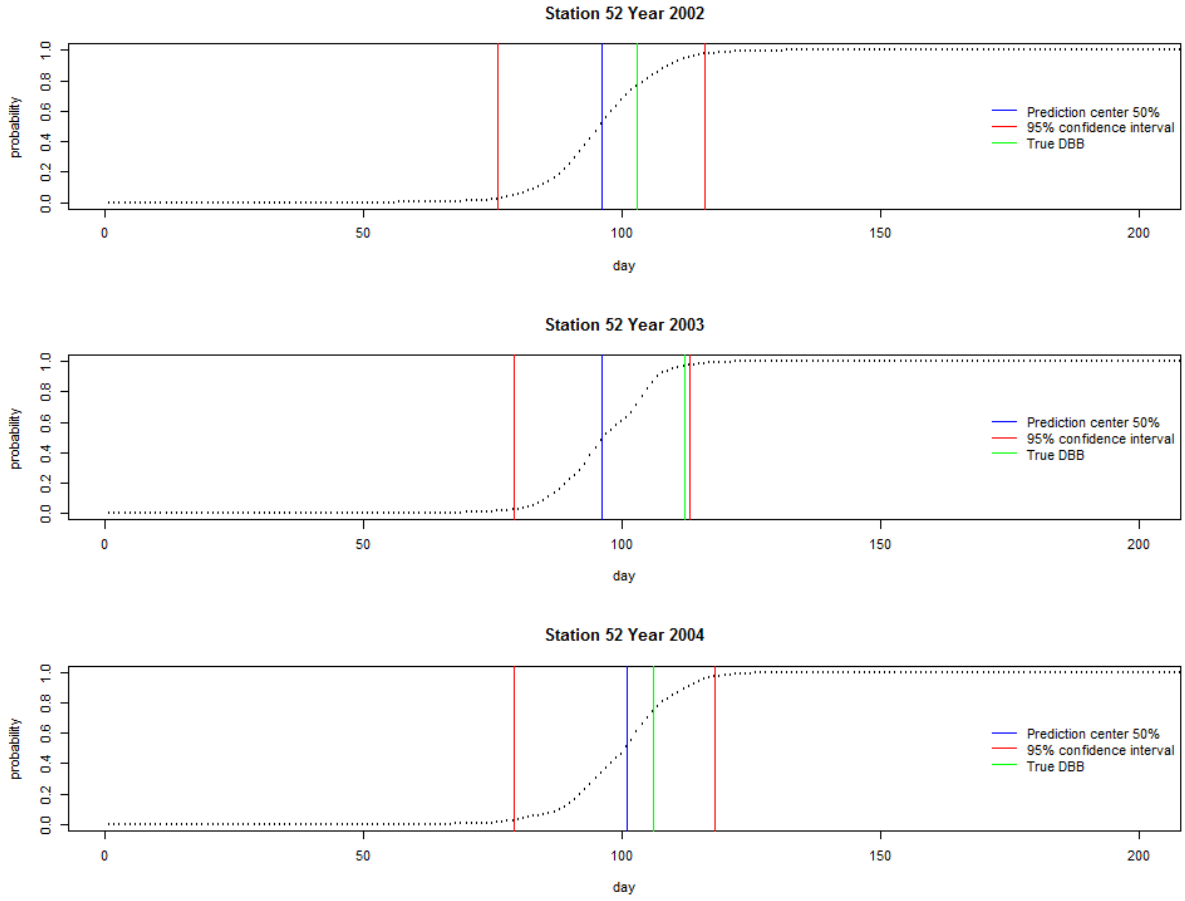

Figure 5.5i Predictions on the UK3 .1se re-estimated and reduced model 1 with the 50% probability marked with blue line. That is the point where the transition probability for budburst is 50%. This is the second variant of PI described in section 5.4. The 95% prediction cover is marked with red lines. The true DBB days are marked with green lines. Note that the transition probability curve is not monotone. Also note that the PI is much smaller than for the first PI variant (figure 5.5h).

6. Conclusions

6.1 Summary

The Markov transition model viewed as a multiple logistic regression model can successfully model the day of bud burst for trees. Setting up the model the regularization method of lasso was used in order to get a manageable and hopefully relevant set of describing covariates. Both models gave a small set of covariates that have the relevant factors discussed both by Song and Kramer. For Finland the constant covariate modeling the total amount of rain the year before stood out. For the UK the covariate modeling day length complemented the covariate modeling the growing degree days.

The Finland model gave very good predictions. Both the MAE and the RMSE for the UK model came very close to their theoretical optimum. Prediction intervals are in both cases too large. For Finland the lower PI limit was shifted downwards and it was not clear how to adjust in general, though the PI could clearly be reduced. For the UK the high variability seemed to create extreme values not captured in prediction. Those values put direct limits on the size of the PI.

6.2 Future work

The approach taken in this bachelor thesis need to be analyzed further to get a larger picture of how the logistic regression modeling of the DBB works in general. The confidence intervals are not obviously defined and needs further understanding. In order to make more accurate prediction on DBB more phenology factors can be considered alongside of bud burst. Another extension of the analysis would be to model a distribution field with Markov properties over time and space.

References

Andersen, E. (1970). Sufficiency and Exponential Families for Discrete Sample Spaces. (*Journal of the American Statistical Association*, Vol. 65, No. 331) 65 (331): 1248–1255.

Becker, R.A.;Wilks, A.R. (2012) Original S code. R version by Ray Brownrigg. Enhancements by Thomas P Minka <tpminka@media.mit.edu> (2012). maps: Draw Geographical Maps. R package version 2.2-8. http://CRAN.R-project.org/package=maps

Brockwell, Peter J.; Davis, Richard A. (1991) Time Series: Theory and Methods (2nd ed.)

Cannel, M.G.R. & Smith, R.I. (1983) Thermal time, chill days and prediction of budburst in *Picea sitchensis*. *Journal of Applied Ecology*, **20**, 951-963.

Casella, G.; Berger, R.L. (2002) Statistical Inference, 2nd ed. Duxbury Press.

Christensen , R. (1990) Log-Linear Models and Logistic Regression

Dykes, L.; Reichel, L. (2012) On the reduction of Tikhonov minimization problems and the construction of regularization matrices - Numerical Algorithms August 2012, 60(4):683-696

Feller, C.; H. Bleiholder, L. Buhr, H. Hack, M. Hess, R. Klose, U. Meier, R. Stauss, T. van den Boom, E. Weber (1995). Phänologische Entwicklungsstadien von Gemüsepflanzen: I. Zwiebel-, Wurzel-, Knollen- und Blattgemüse. *Nachrichtenbl. Deut. Pflanzenschutzd*. **47**: 193–206.

Friedman, J.;Hastie, T.;Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software, 33(1), 1-22. URL http://www.jstatsoft.org/v33/i01/

Gelman, A.; Carlin, J.B.; Stern, H.S.; Rubin, D.B. (2003) Bayesian Data Analysis, 2nd edition. CRC Press.

Hastie, Tibshirani, Friedman (2009) The Elements of Statistical Learning

Climate data from: http://eca.knmi.nl/download/ensembles/ensembles.php

Haylock, M.R., N. Hofstra, A.M.G. Klein Tank, E.J. Klok, P.D. Jones, M. New. (2008): A European daily high-resolution gridded dataset of surface temperature and precipitation. J. Geophys. Res (Atmospheres), 113, D20119, doi:10.1029/2008JD10201"

Kramer, K. (1994) Selecting a model to predict the onset of growth of *Fagus sylvatica*. *Journal of Applied Ecology* 1994, **31**, 172-181

Norris, J.R. (2009) Markov Chains.

PEP725 project (2012) PEP725 Pan European Phenology Data. Data set accessed 2012-09-07 at http://www.zamg.ac.at/pep725/

Picard, Richard; Cook, Dennis (1984) Cross-Validation of Regression Models. *Journal of the American Statistical Association* **79** (387): 575–583.

Rawlings, J.O.;Pantula, S.G.;Dickey, D.A. (2001) Applied Regression Analysis – A Research Tool – Second edition

Song, C. (2010) Stochastic Process Based Regression Modeling of Time-to-event Data, Application to Phenological Data.

Tibshirani, R. (1996) Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society*. Series B (Methodological), Vol. 58, No. 1. (1996), pp. 267-288.

Vegis, A. (1964) Dormacy in higher plants. *Annual Review of Plant Physiology*, **15**, 185-224.

Warnes, G.R. (2012) Includes R source code and/or documentation contributed by: Ben Bolker, Lodewijk Bonebakker, Robert Gentleman, Wolfgang Huber Andy Liaw, Thomas Lumley, Martin Maechler, Arni Magnusson, Steffen Moeller, Marc Schwartz and Bill Venables (2012). gplots: Various R programming tools for plotting data. R package version 2.11.0. http://CRAN.R-project.org/package=gplots

Appendix – Covariates

Constant	Formula	Variable(s)
Chilling days last fall	Threshold values -5, 0,5,10 degrees Celsius. The number of days last fall having minimum temperature below thresholds. Calculated over Oct-Dec	**chillm5, chill0, chill5, chill10**
Last year mean temperature		**mean.year**
Last fall mean temperature	Calculated over Oct-Dec	**mean.fall**
Number of rain days last year divided into periods	Calculated over periods: spring (May-June), summer (July-August), fall (September-December)	**number.spring, number.summer, number.fall**
Total rain last year divided into periods	Calculated over periods: spring (May-June), summer (July-August), fall (September-December)	**total.spring, total.summer, total.fall**
Latitude of station		**latitude**
Height of station	Height for climate and DBB stations	**height.dbb, height.climate**
Continentality	A = max(T_mean_monthly)-min(T_mean_monthly); CI_c = 1.7*A'./sin((lat+10)*pi/180) - 14; CI_g = 1.7*A'./sin(lat*pi/180) - 14; where T_mean_monthly is the average monthly temperature (average over days in month and all years) for each month.	**continentality.cic, continentality.cig**
Varying	**Formula**	**Variable(s)**
Growing degree days GDD	Accumulated temperature above a threshold -2, 0, 5 degrees Celsius from January 1	**gddm2, gdd0, gdd5**
Chilling degree days CDD	Accumulated cold temperature below a threshold -2, 0, 5 degrees Celsius from January 1	**cddm2, cdd0, cdd5**
Growth season	Number of days since the beginning of the growth season. I.e. the first occurrence of four consecutive days with temperature above 5 degrees Celsius	**growthseason**
Frost days	Number of days with frost since the beginning of the growth season. I.e. number of days with temperature below -2 degrees Celsius	**frostdays**
Last frost	Number of days since the last frost beginning from the growth season	**lastfrost**
Day length	Day length in hours	**daylength**
Temperature	Average daily temperature	**temperature**