Using Logistic Regression and Variable Selection to Model Time-To-Event Data

Applications to Tree Phenology and Graduation Time of Engineers

Jesse Burström

Master's thesis
2013:E41

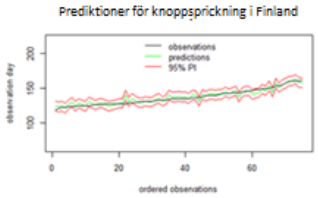
Lund University

Centre for Mathematical Sciences
Mathematical Statistics

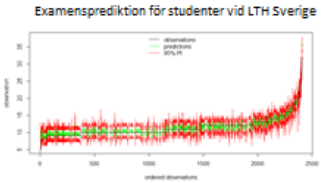
Centrum Scientiarum Mathematicarum

Masters uppsats av Jesse Burström juli 2013
Logistisk Regression för Modellering av Tid-Till-Händelse Data: Tillämpning på Träd Fenologi och Examens Tid för Ingenjörer

När trädens blad knoppar på våren och färgas på hösten är två exempel på tid-till-händelse processer som är påverkade av klimat data. Tid till examen eller till att avbryta studier för ingenjörer är ett annat exempel på tid-till-händelse data, med den ytterligare komplikationen av att ha två möjliga sluttillstånd. I denna masteruppsats vidarutvecklas en modell först presenterad av Song (2010). Tid-till-händelse modellen utökas till att hantera många olika tidsberoende variabler och olika regulariseringsmetoder tillämpas för att hitta statistisk optimala modeller. Klimat variabeln ackumulerad temperatur visar sig kunna förutsäga knoppsprickning med stor precision, se Figur 1, medans tidpunkten för höstlövsfärgningen är mycket svårare att modellera med den givna klimatdatan. Motsvarande prediktioner för examens studenter presenteras i Figur 2. Examensmodellen visar sig vara lik knoppsprickningsmodellen och ackumulerad poäng för studenter spelar samma roll som ackumulerad temperatur för träd. För examensmodellen är det möjligt att klassificera om studenten kommer ta examen eller hoppa av under studiegången termin för termin. Songs modell visar sig vara effektiv och flexibel för modellering av tid-till-händelse data.



Figur 1 Prediktion av knoppsprickning för Björk träd i Finland



Figur 2 Examensprediktion för ingenjörs studenter vid Lunds Tekniska Högskola

Abstract

The day of bud burst (DBB) and leaf senescence are two examples of time-to-event phenological processes influenced by climate factors. Time to graduation or quitting for engineering students is another example of time-to-event data, with the added complication of having multiple possible outcomes, or absorbing states. This master thesis elaborates upon the models presented in Song (2010) “Stochastic Process Based Regression Modeling of Time-to-event Data”. The time-to-event model is extended to use many different covariates, and Lasso regularization techniques are used for variable selection, resulting in compact and statistically relevant models. Models with multiple outcomes are shown to be able to perform classification of students sequentially over time. For the phenological examples, DBB is predicted with an accuracy of a couple of days while leaf senescence proves to be a harder problem, possibly in need of additional climate data not included in this analysis. Overall the model of Song is shown to have great promise and versatility for modeling of time-to-event data.

Acknowledgments

I want to thank Dr Johan Lindström for giving me this interesting master thesis project. I want to thank for all the help and assisting developing the analysis and of course writing this report. I further want to thank Anna Maria Jönsson and Cecilia Olsson for bringing valuable insight into phenological processes.

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 5
2. Data 7
2.1 Phenological data 7
2.2 Examination data 11
3. Model setup 13
3.1 Markov chain transition model as a logistic regression 13
3.2 Regression 14
3.3 Multiple logistic regression 15
3.4 The multinomial case 15
Summary 17
4. Variable selection by regularization 19
4.1 Ridge regression 20
4.2 Lasso 20
4.3 Elastic net 20
4.4 Relaxed Lasso 21
4.5 Cross Validation to select penalty 21
Summary 23
5. Methods for model evaluation 25
5.1 Error estimates 25
5.2 Prediction intervals 25
5.3 The coefficient of determination 26
5.4 The Null model 26
Summary 27
6. Analysis – Phenology 29
6.1 Covariates 29
6.2 Data selection 30
6.3 Parameter estimates 30
6.4 Predictions 36
6.5 Attempts to improve the model 40
Summary 40
7. Analysis – Examination of engineers 41

7.1 Data preprocessing and covariate selection....41
7.2 Program grouping and categorical standardization....42
7.3 Models....42
7.4 Parameter estimates....43
7.5 Predictions & Classification....46
7.6 Simulation....51
Summary....51
8. Conclusions....53
8.1 Future work....53
References....55
Appendix A– Phenological Covariates....57
Appendix B– Student data Covariates....59

1. Introduction

In this Master's thesis a method for analyzing *time-to-event* data is tested and evaluated. Time-to-event analysis is sometimes also called survival analysis. The objective is to model the timing of specific events through a discrete time approximation of the intermediate time steps leading up to the event itself. The model used in this thesis is taken from Song (2010) 'Stochastic Process Based Regression Modeling of Time-to-event Data'.

Two different models for time-to-event data will be analyzed, using one or two distinct endpoints respectively. The single transition model is used to analyze the day of bud burst (DBB) and leaf senescence (when the leaves color at fall) for birch trees in Finland and the United Kingdom. For the dual transition analysis the time to examination or to quitting, for engineers at LTH is modeled.

The first model is used for the phenological states of trees. It is a single state transition model where the system (tree) goes from one state to another (no bud burst to bud burst). The state of bud burst and the state of leaf senescence are well separated in time and are not likely to be influencing each other (more than in the sense there has to be bud burst in order to have leaf senescence) however the possibility is explored. The data used in the analysis do not contain any censored observations. In addition there might be factors in the environment of the tree, perhaps during the beginning of the year, affecting leaf senescence.

The model for engineers is, on the other hand, a dual state transition where the system (student) can go from one state (studying) to one of two (exam or quitting). This system could also be viewed as two separate single transition models, but then the observations have to be separated into those known to have graduated and those that quit. Unfortunately it is not clear how to classify students in the middle of their education. Therefore, a model having both transition cases is constructed. It is shown that the multiple transition models easily generalize to k-states, in an extension of Song's model.

The model also uses the application of many (hundreds of) covariates in the analysis. They are, through Lasso regularization (Tibshirani, 1996) and grouping (in the case of engineers the 14 different programs are grouped into 3 clusters), reduced to statistically significant depending covariates. The full model described by all the covariates (and their interactions) is often too complex leading to over fitting. The full model is therefore not a good predictor on new data. Cross validation (CV) and regularization is implemented to find covariates that best describe the given data set.

In the following sections the analysis of the phenological states will implement hundreds of climate data derived covariates in order to find models that best describe the data (DBB and leaf senescence). It is shown that the growing degree days covariate analyzed by Song is effective for predicting DBB. The model for the leaf senescence is however not successful, given climate data at hand. One possibility would be to incorporate sun exposure time as a

covariate for the leaf senescence. The analysis of the engineer's path to graduation or quitting is modeled both as single transition models and a dual transition model. The single transition models for the examination and quitting prove to have similar structure as the DBB models for the Finland and UK data. The dual transition model for the engineers is used for classification of student's probability to examination and quitting. It is shown that the classification can be made on censured data i.e. on students at various stages in their education.

2. Data

All analysis in this thesis was done using the open source computing environment R (R core team, 2012). The R-package *glmnet* (Friedman, 2010) provides tools for Lasso analysis of logistic (and other) regression. Specifically version 1.9-3 (March 2013) which supports regularization without an intercept, was used.

2.1 Phenological data

During one season/year a tree goes through a series of phenological stages as budburst and leaf senescence (LS). Leaf senescence is when the leaves change color during the fall to save minerals before falling off. There are many deterministic models that try to identify dependent factors concerning DBB and senescence. For the DBB an important factor is the accumulated temperature described and analyzed by Song (2010). Other factors affecting DBB include chilling degree days (Vegis, 1964) & (Cannell & Smith, 1983). Other models propose the amount of daylight as a key triggering factor for budburst (Kramer, 1994). The idea of chilling or daylight is to prevent early DBB due to warm winter that then risks damage in the case of cold spring months. Further the amount of rain during the year before could affect budburst; stress due to drought could affect the winter rest needed. Therefore climate/weather during different time periods over the previous year are used as covariates. The amount of rain along with number of rain days over these periods are used as constant covariates. For senescence, the covariates decreasing temperatures at fall, possibly in combination with day length is used (Delpierre, 2009).

Phenology data was taken from the PEP725 project (2012). The tree analyzed is birch (lat. Betula Pendula). Two phenological phases are studied; the day of bud burst (DBB) as event 11 in the BBCH scale (Feller, 1995), which is when the first leaf has unfolded; and event 94 leaf senescence, which is when 50% of the leaves have colored during the autumn.

Phenology data for trees have been collected irregularly across Europe. In this thesis the climate data is furthermore interpolated from a regular grid. Data from Finland and the United Kingdom are analyzed.

Climate data on a 0,5 x 0,5 degree grid was obtained from Haylock (2008). Phenological observations are assigned to the closest grid cell. For Finland the grid resolution of the climate data allows for a unique association of each phenology stations. In the case of the UK the number of phenology stations is much larger than the number of climate stations, and many stations have the same climate data despite a substantial spread in DBB within each climate grid cell. See table 2.1 and figures 2.1a-b.

Coastal data are dropped due to lack of climate data. Temperature and precipitation data is collected from the closest climate cell relative the phenology station. The climate data

consists of daily average, minimum, maximum temperatures, as well as daily precipitation (amount of rainfall).

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

Finland

The number of DBB observation sites in Finland is 29 and the number of observation sites for leaf senescence is 31. Figure 2.1c shows the distribution of the observation sites and the choice of training/validation sets for DBB data. The distribution of the observation sites is similar for the leaf senescence data. In figure 2.1d the distribution of observations over the years is shown for each observation site in the case of DBB. A similar distribution holds for leaf senescence data. For the Finland data 2/3 of the stations are selected for the training set and the remaining 1/3 for the validation set.

United Kingdom

For the UK the number of DBB observation sites is 3169 and the number of observation sites for leaf senescence is 2364. Figure 2.1c shows the distribution of the observation sites and the choice of training/validation sets for DBB data. The distribution of the observation sites is similar for the leaf senescence data. In figure 2.1e the distribution of DBB observations over the years is shown. A similar distribution holds for leaf senescence data for the UK but only over the years 1999-2005, i.e. no data for senescence exists before 1999. Due to computer memory limitations only 1/3 for DBB and 1/4 for leaf senescence of the UK data are used for the training set.

Table 2.1 Data for Finland and the UK describing the distribution of observation dates and stations.

	Years	Number of phenology stations	Number of climate stations	Median observation	min/max observation	Number of observations
FI - DBB	1997-2005	29	29	138	119/172	240
UK - DBB	1972-2005	3169	418	103	32/150	6637
FI - LS	1998-2011	31	31	252	210/290	290
UK - LS	1999-2005	2364	418	297	223/362	4450

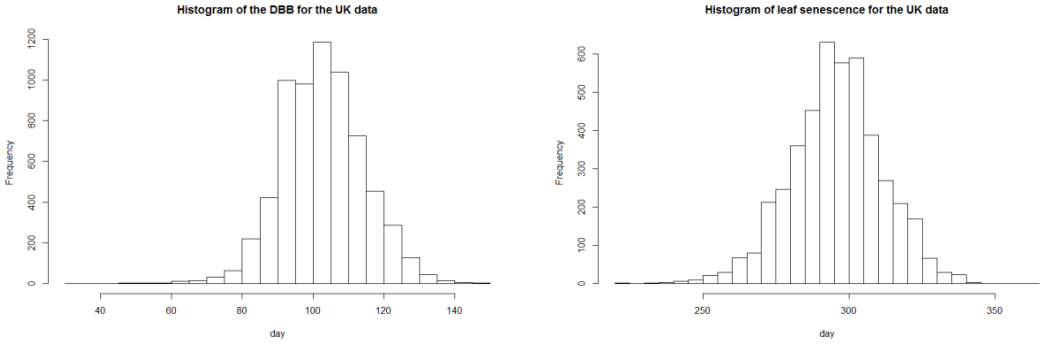


Figure 2.1a Histogram for the UK data showing the distribution of observation dates.

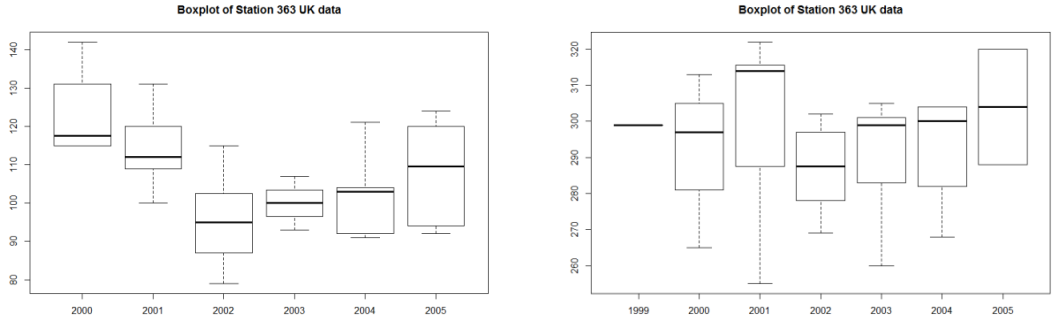


Figure 2.1b Boxplot of weather station 363 UK data showing the variability of observations within a year. To the left is DBB data having 3-19 observations. To the right is leaf senescence data having 1-15 observations.

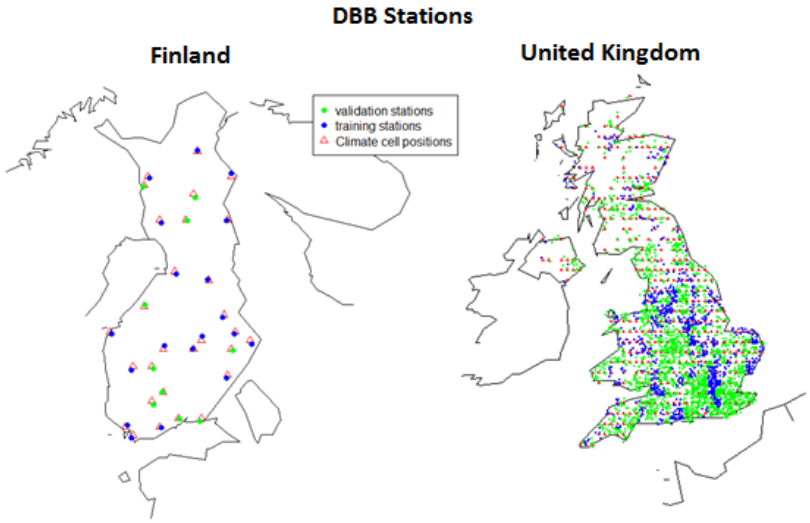


Figure 2.1c Phenology stations for the DBB data with the training set as well as remaining validation set. The climate grid locations closest to each phenology stations are shown. The associated climate stations from Haylock (2008) are taken from a 0.5 x 0.5 degree grid

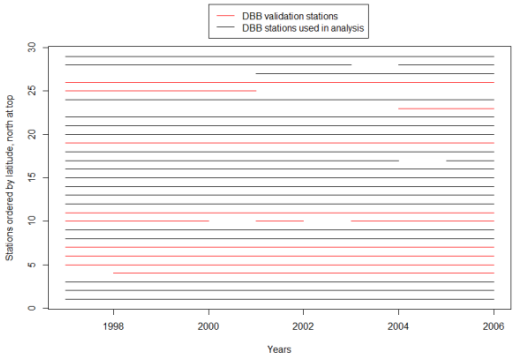


Figure 2.1d The Finland DBB observation for each year and phenology station.

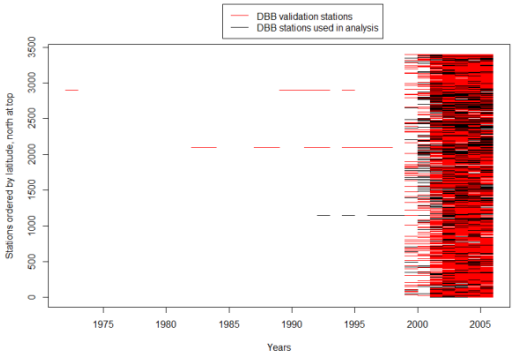


Figure 2.1e UK phenology stations and their observations over the years. Almost all observations are from 1999-2005

2.2 Examination data

The data for the analysis of engineer examination is taken from LTH (faculty of engineering at Lund University (LU)) and consists of 18434 students at 14 different programs during the years 1993-2012. Each student is represented during each semester from registration at a program to either graduating or quitting. Some students may be represented multiple times with different identification numbers due to change of program. However it is not possible to distinguish these cases from other cases. Therefore a portion of the quitting students only change program and are registered cases more than once. This affects the true number of quitting students. Of the data the following information is used:

- Program – Engineering program identification letter
- Identification – Individual identification number for each registration
- Female – Indicates a female student
- Class – Registration year and semester.
- Semester – The current semester.
- Status – Activity status for the specific semester: Quitted, Graduated, Inactive, Foreign studies, Study break
- Points – Registered points for the current semester in the LADOK system of Lund.
- Study semester – The study semester number ordered from start for each student.

The engineering programs are identified by letters, with some programs started after the first year for which the data set contains students, see table 2.2. The number of student registrations per year divided into male and female is shown in figure 2.2a. Figure 2.2b shows the distribution of how many students take exam and quit per year in the data set.

Table 2.2 Engineering programs in data set that span 1993-2012.

Programs	Years
D E F K L M V	1993-2012
I W	1998-2012
B C	2001-2012
P	2002-2012
G N	2003-2012

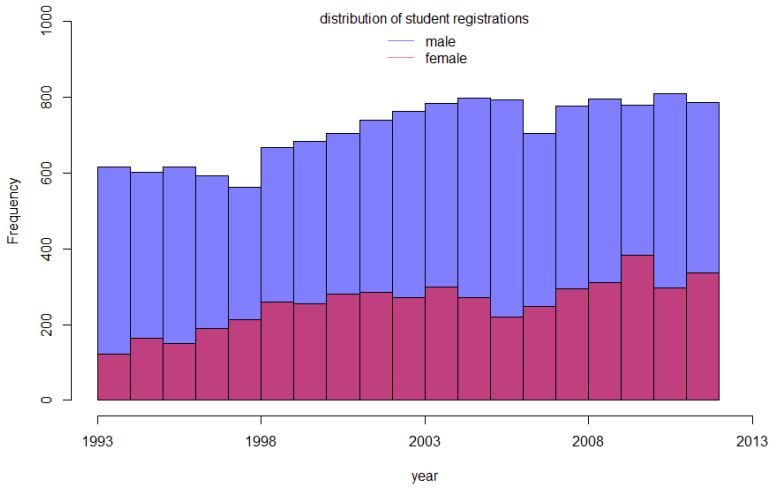


Figure 2.2a Histogram of student registrations from first semester Fall-1993 to Fall-2012.

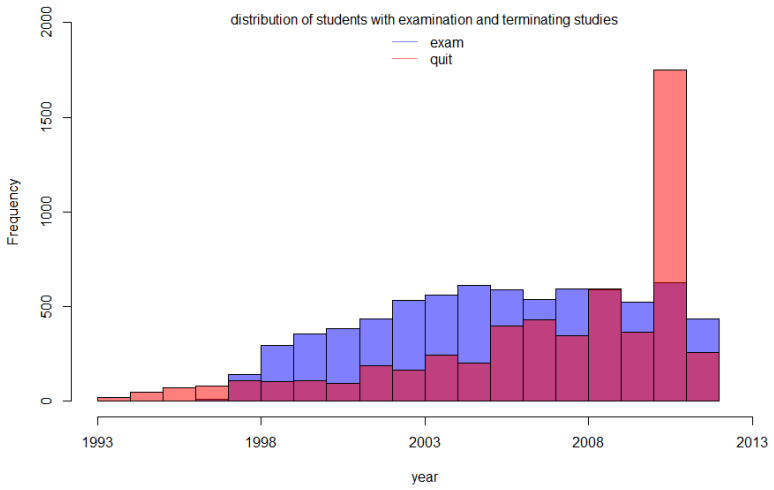


Figure 2.2b Distribution of the number of students taking out exam and terminating studies. The first examinations come roughly 5 years after the first students started their program. Note the increase in quitting around year 2011, it is due to a large database clean out of inactive students.

3. Model setup

In this thesis *time-to-event* data is analyzed using the stochastic regression model proposed by Song (2010, chapter 2.2). Formally time-to-event is defined as a sequential process over time $t = 0,1,...$ going through state transition. The states are described by an indicator variable $Y_{i,t} \in \{0,1\}$ where the index i represents each individual or case studied see figure 3.1. In case of budburst 0 stands for 'budburst has not occurred' and 1 stands for 'budburst has occurred'. Correspondingly for the engineer's data the indicator variable has three states {1,2,3} where 1 stands for 'studying', 2 stands for 'exam' and 3 stands for 'quitting'. Now let T_i be the time to event for observation i then:

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

In addition we assume there are time dependent vectors of covariates, $X_{i,t}$ which affect T_i.

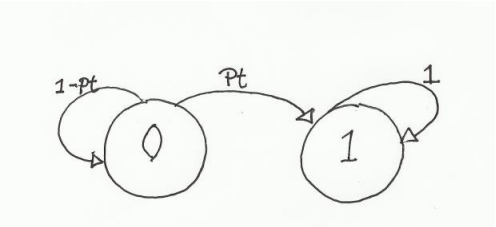


Figure 3.1 Two state Markov transition model. The Markov chain (Norris, 2009) is in the 'zero' state and remains there at each time point t with probability $1 - P_t$. Consequently the chain changes state with probability P_t at each time point t. When the Markov chain has changed state to the 'one' state it will remain there with probability 1 and has thus been absorbed by the 'one' state.

3.1 Markov chain transition model as a logistic regression

The conditional probability of transition at time point t_i given covariates is:

$$\mathrm{P}(T_i = t_i \mid X_i) = \mathrm{P}(Y_{i,0} = Y_{i,1} = \ldots = Y_{i,t_i-1} = 0, Y_{i,t_i} = Y_{i,t_i+1} = \ldots = 1 \mid X_i)$$

This states that the probability of transition at t_i is the probability of having the chain of indicator events up to and after the transition. Since we are interested in the time-to-event conditionally on the covariates $X_{i,t}$ we write the probability for the chain of events $Y_{i,0:t}$ given $X_{i,0:t}$ as:

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

It is now easy to show, due to the sequential nature of the time-to-event model, that the probabilities on the right hand side of the above equation only depend on $Y_{i,s-1}$ (Song, 2010, (2.5)). The model simplifies to:

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

And we get:

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

By the construction and assumptions of causality (see figure 3.1 for illustration of the Markov model) it is clear that the covariates can only influence transitions up to their present time, thus X_i should be replaced by $X_{i,0:s}$ in the above equations.

We now assume that the transition probabilities can be written as:

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

Then the probability of having transition at time $T = t$ is:

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

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

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

3.2 Regression

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

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

The explanatory variables can have many forms: continuous, constant (the intercept being the default constant equal to one) or categorical. However, the response Y must be continuous in order for least squares to make sense. So in the case of a categorical response as the 0/1 state coding of Y in the time-to-event transition model, linear regression is not applicable

3.3 Multiple logistic regression

When the response is binary, such as $Y_i \in \{0,1\}$ in the time-to-event Markov chain (see figure 3.1), Y_i can be modeled as $Bin(1,\mathrm{p}) = Bernoulli(\mathrm{p})$. Trying to set up a regression modeling the success probabilities p_i as,

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

fails since $p_i \in [0,1]$ and we cannot guarantee that $p_i = X_{.i}\beta \in [0,1]$. In order to overcome this problem one can use a *link-function* such as the *logistic function*:

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

This transformation maps the probabilities to $\mathbb{R}$.

The single transition model will for each observation fit a value on the real line and the transformation, using the inverse logistic function:

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

converts the values back to probabilities. So the model:

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

inverted with the logistic function gives

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

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

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

We have the response y_i for the observation vector $X_{i.}$. The maximization of the likelihood gives the estimate of β and the transition probabilities are given by $P_i(\mathrm{X}_{i.}\,\beta)$. Note that the maximization is done over all observations T_i represented by the corresponding sequences $y_{i;o:t_i}$.

The resulting model has the properties of a general linear model (GLM) among which is that the coefficient estimations are consistent maximum likelihood estimates. Simulation studies (Song, 2010, ch 3.4-3.5) confirms the consistency of the estimates.

3.4 The multinomial case

For the examination data, the two separate absorbing states results in the use of a multinomial analogue to the binomial distribution (figure 3.4). The derivation in chapter 3.1 holds, with the modification of (1) allowing for multiple transitions.

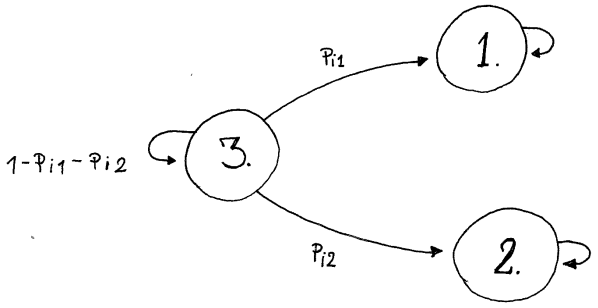


Figure 3.4 The multinomial Markov chain. For each observation i there is a probability p_{i1} of going to the absorbing state 1 and a probability p_{i2} of going to the absorbing state 2. Consequently there is a probability $p_{i3} = 1 - p_{i1} - p_{i2}$ of staying in the state 3. The observation i is part of a sequence making up the Markov time chain.

The distribution function for k categories is:

$$f(\mathrm{y}_{i1},\ldots,\mathrm{y}_{ik},\mathrm{n},p_{i1},\ldots,p_{ik}) = \begin{cases} \dfrac{n!}{y_{i1}!\cdots y_{ik}!}\, p_{i1}^{y_1}\cdots p_{ik}^{y_k}, & \text{when} \sum_{j=1}^{k} y_{ij} = n \\ 0 & \text{otherwise} \end{cases}, \text{ with } n = 1 \quad (2)$$

The corresponding binomial case has k=2 categories, one for success and one for fail. Here we have $k = 3$, $\mathrm{y}_{i\cdot}$ is one of $(1,0,0),(0,1,0),(0,0,1)$ so that only one of the response states is possible for each observation i. The probabilities are for the transition to the 'exam' state, the transition to the 'quit' state and of remaining in studies which is given by the probabilities of the two transition states since the total probability sums to one i.e.

$$\sum_{j=1}^{k} p_{ij} = 1, \ \forall i$$

Since there are more than two states in the logistic regression model one state has to be chosen as a reference state. Compared to the logistic function, in the divisor is the failing case, $1-p$, for the binomial logistic regression and the dividend, p, the success. With more than two categories the divisor is chosen to be the 'no event' probability or simply 1 minus the sum of the probabilities of the absorbing states. Since we have $k-1$ probability sequences for the different transition states the logistic regressions becomes:

$$\ln\left(\frac{p_{i1}}{p_{ik}}\right) = X_{i\cdot}\beta_1, \ \ldots\ , \ \ln\left(\frac{p_{i(k-1)}}{p_{ik}}\right) = X_{i\cdot}\beta_{k-1}$$

Exponentiation and solving for the probabilities gives:

$$p_{i1} = p_{ik}\exp\left(X_{i\cdot}\beta_1\right),\ldots,p_{i(k-1)} = p_{ik}\exp\left(X_{i\cdot}\beta_{k-1}\right)$$

Since the probabilities must sum to 1 we get:

$$p_{il} = \frac{\exp\left(X_{i\cdot}\beta_l\right)}{1+\sum_{j=1}^{k-1}\exp\left(X_{i\cdot}\beta_j\right)},\ l = 1,\ldots,k-1, \text{ and } p_{ik} = \frac{1}{1+\sum_{j=1}^{k-1}\exp\left(X_{i\cdot}\beta_j\right)} \quad (3)$$

The connected equations described by (3) are then solved using ML over the multinomial distribution (2) giving the multinomial log likelihood,

$$\log\prod_{i=1}^{n} f(\mathrm{y}_{i1},\mathrm{y}_{i2},\mathrm{y}_{i3},1,p_{i1},p_{i2},p_{i3}) = \log\prod_{i=1}^{n}\frac{1!}{y_{i1}!\,y_{i2}!\,y_{i3}!}\,p_{i1}^{y_{i1}}p_{i2}^{y_{i2}}p_{i3}^{y_{i3}} =$$

$$\sum_{i=1}^{n}\log\left(\left(\frac{\exp\left(X_{i\cdot}\beta_1\right)}{1+\sum_{j=1}^{2}\exp\left(X_{i\cdot}\beta_j\right)}\right)^{y_{i1}}\left(\frac{\exp\left(X_{i\cdot}\beta_2\right)}{1+\sum_{j=1}^{2}\exp\left(X_{i\cdot}\beta_j\right)}\right)^{y_{i2}}\left(\frac{1}{1+\sum_{j=1}^{2}\exp\left(X_{i\cdot}\beta_j\right)}\right)^{y_{i3}}\right)=$$

$$\sum_{i=1}^{n}\log\left(\frac{\exp\left(y_{i1}X_{i\cdot}\beta_1\right)\exp\left(y_{i2}X_{i\cdot}\beta_2\right)}{1+\sum_{j=1}^{2}\exp\left(X_{i\cdot}\beta_j\right)}\right)=\sum_{i=1}^{n}y_{i1}X_{i\cdot}\beta_1+y_{i2}X_{i\cdot}\beta_2-\log\left(1+\sum_{j=1}^{2}\exp\left(X_{i\cdot}\beta_j\right)\right)\ \text{when}\sum_{j=1}^{3}y_{ij}=1$$

Here the 3 state is the ‘no transition’ state and the states 1 and 2 the two absorbing transition states. The probability of transition within each observation sequence as calculated in chapter 3.1 is similar for the multinomial model. In this thesis the two coefficient sequences β_1,β_2 represent the same covariates (the ‘grouped’ option in glmnet). It is possible to have different covariates for the two transition states (the corresponding ‘ungrouped’ option). Intuitively it is though plausible to use the grouped model since often a good covariate for examination is bad for quitting and vice versa. For example: points is good for examination and bad for quitting, consequently the probabilities will go up for examination and held down for quitting.

Summary

As shown the time-to-event data has a stochastic process based regression representation (Song, 2010) in the form of a Markov chain where dependence on covariates can be estimated using logistic regression. One key aspect of the regression is that each observation event is made up of a sequence of time points coding the chain. For the phenological data the sequence consists of all the days from January 1 up to the phenological event. For the engineer’s data the sequences are the semesters autumn/spring up to the event exam/quit. The logistic regression itself makes no distinction between the sequence parts of the observations. For example: day 15 in observation i is treated the same as day 123 in observation j. What make the two elements different is the covariates. The logistic regression gives the same answer for any elements that have the same covariates. This makes the time constraints on the covariates more apparent since the covariates bear all information of the time dependence throughout the logistic regression. However the binomial distribution of the observations defines the relative probabilities within an observation set. For predictions there will be a difference since the probabilities are sequentially summed within each observation sequence. The logistic regression finds the coefficients that provide the best overall fit to all observations. Therefore it is (again) essential that there are mathematical structures in the covariates that capture the direction from time point 1 up to the transition point in order for the regression to generate models with good predictive properties. It is, however, a model assumption that the covariates influence the time-to-event. It is also why the accumulated version of any given affecting covariate, in this model, can make a good predictor, since the accumulation defines a time direction within each observation sequence.

4. Variable selection by regularization

In a regression model the problem of finding good predictors and discarding irrelevant ones becomes increasingly hard as the number of possible covariates increases. The regression will use as many variables as possible to minimize the sample error, often leading to over fitting. This can be resolved by penalizing models with many parameters, allowing a compromise between model complexity and fit. Different forms of information criteria have been developed to help choose between models with differing sets of covariates. The two most common are the Akaike information criteria (AIC) (Akaike, 1974) and the Bayesian information criterion (BIC) (Schwarz, 1978). The general form of the AIC is:

$$AIC = 2k - 2\ln(L)$$

where k is the number of the parameters in the model and L is maximum likelihood of the fit. A low value of the AIC indicates a better model. This measure is only relative, allowing us to compare different models; it does not say if a particular model provides a good fit to data. Typically one uses AIC or BIC to choose between nested or related models already known to be reasonable predictors.

Another slightly subtle form of over fitting occurs when validating several models. Say that the given data set is divided into two groups, a training set used to estimate model parameters, and a validation set for evaluation of the models. Each model will then be given a validation error and it is natural to choose the model with the smallest validation error as the best. However, here the validation error is explicitly minimized and therefore biased towards the validation set. To overcome this problem the training set is further divided into groups so that *cross validation* (CV) can be applied (Picard, 1984). The idea is to successively hold out one group when estimating parameters and then use that group for validation. The number of cross validation groups is often set to 10 and the resulting 10 validation errors are weighted to form a final validation error. The model with the lowest combined error is then chosen and can thereafter be independently validated on a remaining validation set.

It is often desirable to automatically find a good reduced model when the number of covariates is very large. Using information criteria or cross validation over all possible subsets is only computationally feasible for a limited number (about 10) of covariates. The idea of penalizing the model for having many parameters as done for the AIC and BIC information criterions can be applied more directly in the regression estimation itself. Adding a term involving the magnitude of the regression coefficients scaled by a factor λ to the log-likelihood forces some coefficients to become small. The model selection process can then focus on coefficients above a certain threshold. The overall process involves selecting the λ parameter, often through CV, and scaled by a threshold. When applying regularization it is important to standardize the covariates so that they have the same scale. This is done by centering, that is subtracting the mean, and scaling, dividing by the standard error. If the model is without an intercept one can only scale. Constant covariates are not scaled and categorical covariates are not standardized at all.

4.1 Ridge regression

One common method of regularization is ridge regression, or Tikhonov regularization (Tikhonov, 1943). As penalty the sum of the squared coefficients is added. For linear regression the minimization becomes:

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

The intercept is not penalized in order to avoid dependence on the origin of **y**. Although this expression is the sum of the squared errors (with an extra penalizing term) it can also be seen as the maximization of the log likelihood of joint Gaussian distribution where the penalizing factor becomes a Bayesian prior, on the coefficients,

$$p(\beta) = \exp\left(-\lambda \sum_{j=1}^{p} \beta_j^2 \right) \text{ or } \beta_j \sim N(0, 1/(2\lambda))$$

4.2 Lasso

Tikhonov regularization accomplishes the goal of reducing the number of covariates, if a threshold is chosen. The Lasso (Tibshirani, 1996) overcomes the problems of choosing coefficient threshold by using the sum of the absolute values of the coefficients in the penalizing term. Contrary to ridge regression the regularization in Lasso is not smooth at the origin, forcing coefficients to zero as lambda increases.

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

The Lasso has no closed form solution, but it constitutes a quadratic programming problem, related to convex optimization (Hastie, 2009, ch 3.4.2), and efficient algorithms exists. The corresponding Bayesian prior distribution is $p(\beta) = \exp(-\lambda|\beta|)$, a Laplace distribution (Tibshirani, 1996, ch 8). For the logistic regression the Lasso estimation is:

$$\hat{\beta}^{lasso} = \arg\min_{\beta} \left\{ -\sum_{i=1}^{N} \left(y_i X_{i.}\beta - \ln(1 + \exp(X_{i.}\beta)) \right) + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

4.3 Elastic net

When Lasso is used for variable selection it still has a few shortcomings. For strongly correlated covariates Lasso tends to select only a few of the correlated covariates, and it has been empirically observed that ridge regression provides better predictive performance for highly correlated predictors. Further, if the number of predictors is larger than the number of observations, $p > n$, Lasso select at most n variables. To alleviate these limitations a combination of ridge and Lasso regularization can be used. The property of the ridge regression to group strongly correlated covariates is combined with Lasso's ability to do variable selection by forcing coefficients to zero. Moreover the combined regularization allows for more predictors than observations. The combined regularization prior is known as the elastic net (Zou, 2004) and consists of a by-linear combination of priors,

$$p(\beta) = \exp(-\lambda(\alpha \mid \beta \mid_2^2 + (1-\alpha) \mid \beta \mid_1)), 0 \le \alpha \le 1$$

Where $\alpha = 0$ corresponds to Lasso and $\alpha = 1$ to ridge regression.
It has also been shown (Zou, 2004) that an efficient algorithm exists (LARS-EN) to solve the elastic net optimization; empirically elastic net outperforms both Lasso and ridge regression.

4.4 Relaxed Lasso

For Lasso and elastic net the penalty parameters α, λ are often chosen through CV. The cross validation procedure often chooses more covariates than needed. These covariates, called noise features, are the result of how the Lasso algorithm shrinks the coefficients and mainly occur for high dimensional data. An alternative algorithm, the *relaxed Lasso* (Meinshausen, 2006), uses a variation of the Lasso algorithm where each model produced for a given λ parameter, is re-estimated through a new Lasso regularization with an extra parameter $\phi \in [0,1]$ relaxing the λ penalty. The re-estimation uses only the covariates selected by the first Lasso regularization but with a relaxed penalty $\phi\lambda$,

$$\hat{\beta}^{\lambda,\phi} = \arg\min_{\beta} n^{-1} \sum_{i=1}^{n} (Y_i - X_i^T \{\beta \cdot 1_{\mathfrak{M}_\lambda}\})^2 + \phi\lambda \parallel \beta \parallel_1$$

Here $1_{\mathfrak{M}_\lambda}$ is the indicator function on the set of variables $\mathfrak{M}_\lambda \subseteq \{1,...,p\}$ selected in the first Lasso so that for all $k \in \{1,...,p\}$

$$\{\beta \cdot 1_{\mathfrak{M}_\lambda}\} = \begin{cases} 0, & k \notin \mathfrak{M}_\lambda \\ \beta_k, & k \in \mathfrak{M}_\lambda \end{cases}$$

Intuitively Lasso shrinks the parameters in the selection process. The resulting model will therefore have smaller coefficients than the corresponding GLM estimate. Due to this shrinkage the Lasso model can have a number of noise features, variables with small coefficients that over fit the data. Relaxing the penalty will grow the coefficients, punishing (by increasing the CV-error) models with too many parameters.

For $\phi = 1$ we get the ordinary Lasso solution but for smaller values the combined penalty is reduced. At $\phi = 0$, which is a degenerate limiting case of the relaxed Lasso, each model is simply re-estimated without any penalty corresponding to an ordinary regression. It has been shown that the corresponding subset of models favor sparser models at the minimum, with equal or improved predictive power (Meinshausen, 2006). For the purpose of this thesis the relaxed Lasso with parameter $\phi = 0$ is used.

4.5 Cross Validation to select penalty

Regularization is essential when having a large set of covariates. The goal is to find a model that is as compact as possible while providing maximal predictive power. If the model contains too many parameters the fit to the training set may be good, but at the risk of over fitting, making the model poor on new data. When applying regularization a range of values for the parameter λ is fitted as penalty. This gives a range of models with different model sizes. As the penalty grows for the Lasso regularization the resulting model is typically smaller since coefficients are forced to zero.

In order to get an unbiased cross validation all parts of the training set are used by dividing the training set into k parts each in turn used for cross validation. The total error is then weighted. This can be an important step especially if the training set is small (compared to the *model size*) since the validation can (strongly) vary due to *outliers*. Essentially cross validation is a much more stable estimate than using the whole training set for model estimation (the effect of outliers are averaged).

By an *outlier* is meant an observation that is badly modeled in the sense that it gives an unusually large error. Sometimes it can be due to measurement errors but it can also happen that the data includes some exceptional cases. The resolution could be to either remove the outlier (in the case of error in data measurement) or to extend the model in an attempt at capturing the exceptions.

The regularized models will have a CV standard error as seen in figure 4.5. The model with the smallest CV error is chosen to be the best. However within one standard error from the optimal model there are often models with smaller model size. One can argue that models within one standard error of the optimal model cannot be statistically distinguished. It is therefore recommended to pick the model with smallest model size and smallest error within one standard error of the optimal model. This selection can be used both for models generated by the Lasso (elastic net) regularization and the corresponding models from the relaxed Lasso.

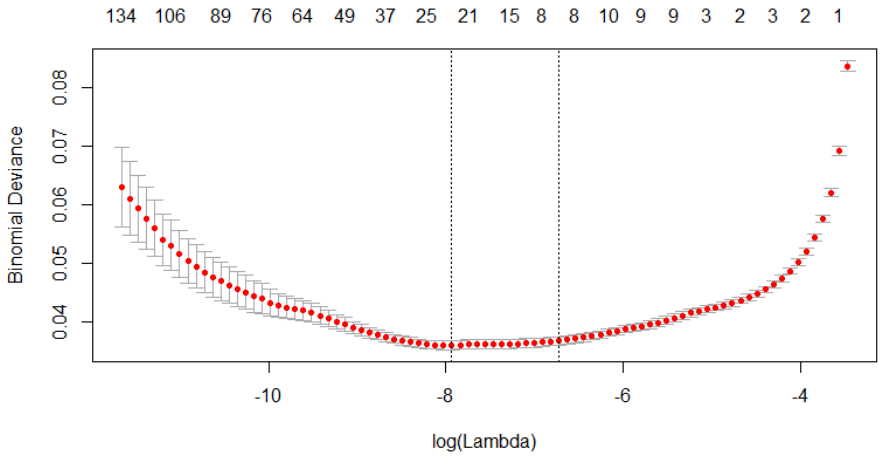


Figure 4.5 The cross validation give rize to a sequence of values of lambda having different CV-error as well as CV-variance.

Summary

When choosing a model through regularization several criteria are often set, some are:

- Ability to handle high dimensional data where possibly $p \gg n$
- Computational efficiency without compromising model quality
- Consistency in variable selection
- Optimal predictive power

These constraints seem to be met by a combination of covariate selection (prior to the model estimation), the elastic net regularization, relaxed Lasso and one standard error approximation of the model. In this thesis the minimum model and the smallest one standard error model for both the elastic net and the relaxed Lasso are compared for a range of elastic net α - values.

5. Methods for model evaluation

When a model has been decided on (i.e. estimated) it needs to be evaluated. The standard approach is to test the model on data not used in the estimation process and measure the misalignment or error. The data to be modeled is therefore divided into two parts: The first is called the *training set* and usually makes up most of the data (here 2/3). The second part is called the *validation set* and is the remainder of the data (here 1/3). The training set is used to estimate the parameters of the model, possibly using CV. The model found using training data can then be evaluated using the validation set.

5.1 Error estimates

As a part of the validation process there needs to be some measure of model fit.
The most common error estimate used for cross validation is the root mean square error (RMSE).

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}$$

Another option is the mean absolute error (MAE).

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i|$$

Due to the square, RMSE is more sensitive to large outliers.

5.2 Prediction intervals

The error estimates for a given model produces a rough indication of the variance of the estimation. In fact the MSE is related to the variance by:

$$MSE(\hat{\theta}) = E\left[(\hat{\theta}-\theta)^2\right] = Var(\hat{\theta}) + \left(Bias(\hat{\theta},\theta)\right)^2$$

This is also sometimes referred to as the *variance bias tradeoff*. As the model size increases the variance will typically increase and the bias will decrease. The error is minimized for a certain model complexity. For a model with small bias the MSE is an estimation of the variance. For the 10-fold CV taking the square root of the MSE and multiplying with 1.96 to give an approximate 95% prediction interval:

$$PI_{0.95} = \left[\hat{\theta} - 1.96\sqrt{MSE}, \hat{\theta} + 1.96\sqrt{MSE}\right],$$

since the MSE is approximately normal for a large training set. The prediction interval length is then $2 \cdot 1.96 \cdot RMSE$. For the GLM estimation the MSE is normally calculated with respect to the probabilities but in this thesis the MSE with respect to the predicted transition time is the error used for CV model selection. For the dual models the combined prediction errors are used for the model selection.

5.3 The coefficient of determination

There is a measure of the statistically explanatory power of the underlying data for each of the models called the *coefficient of determination* R^2. It measures the amount of variation in the data explained by the model and is defined as,

$$R^2 = 1 - \frac{SS_{err}}{SS_{tot}} = 1 - \frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{N}(\bar{y} - y_i)^2}$$

If the error estimate is close to zero R^2 will be close to one, which is the optimal value. However, if the estimates predict the true values worse than the observation mean, R^2 will be negative. In this thesis R^2 is used to evaluate the fit to the data.

In a linear model with repeated observations

$$y_{ij} = X_i\beta + \varepsilon_{ij}$$

$y_{\cdot j}$ has the same covariates for all j. In this case it is possible to split the variance in two parts (Montgomery, 2001),

$$V(y_{ij} - X_i\beta) = V(y_{ij} - \bar{y}_i) + V(\bar{y}_i - X_i\beta) \qquad (4)$$

It is clear that the total variance will always be at least as large as the first term in (4). Assuming perfect predictions the second term of (4) will be zero. The largest possible R^2 value will therefore be,

$$R^2_{\max} = 1 - \frac{V(y_{ij} - \bar{y}_i)}{V(y_{ij} - \bar{\bar{y}})}$$

As an example; if the climate covariates for a set of phenological observations have low geographical resolution, as for the UK data, the geographically close observations will get the same covariates, the first term of (4) will be positive and $R^2_{\max} < 1$.

5.4 The Null model

As a reference model of the efficiency of a specific model one can use what is sometimes referred to as *the null model*. By the construction of the coefficient of determination one sees that a simple model containing no covariates at all but has as predictor the mean of the observations in the training set would give a R^2 of roughly zero (depending on the statistical variability between the training set and the validation set). Thus a model using the mean of the observations as prediction is introduced as our null model. The error estimates are calculated in the same way over the validation set using the mean of the observations in the training set. The confidence intervals are formed as above using the RMSE of the training data.

Summary

In order to assess the quality of a given model the model's ability to do prediction on new data is evaluated. As it turns out the RMSE of the cross validation gives a rough estimation of the prediction interval for the model. One sees that the estimated transitions actually are point estimates making the PI approximation plausible. The process of assessing the quality of a model is hierarchical as the evaluation of a model not only takes into account the performance on data but also against the reference null model which opens up the possibility of comparing models. When comparing models it is preferred that the dataset is identical and even the training and validation set identical or at least the validation set identical. Therefore, in practice, the comparison is only general.

6. Analysis – Phenology

This analysis sets out to investigate which climate factors best predict the day of bud burst (DBB) for birch (lat. *Betula pendula*) and the day when 50% leaf senescence for 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 which are considered in different combinations.

The analysis for DBB is inspired by Song (2010) who analyzed the DBB of different tree species in Canada with respect to growing degree days (GDD). This analysis expands on the work of Song by investigating the effects of many factors, and how they best can be used to construct a forecast model for predicting DBB and senescence.

For leaf senescence the work of Delpierre et al. (2009) is considered as inspiration for covariate selection. Variations on the idea of decreasing degree days after the summer solstice are explored. However the prediction of leaf senescence proves to be exceptionally difficult and no conclusive explanatory factor is found. Instead attempts to identify covariates are explored and some suggestions are made regarding additional climate data that could possibly be used to form good predictors.

The aim of this analysis is to take a statistical approach to modeling of bud burst and leaf senescence. 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 aid in finding other relevant biological factors linked to the DBB and even so could be used with climate models for prediction. Furthermore a range of statistical modeling techniques are compared to assess the consistency of the derived models and the obtained relevant covariates.

6.1 Covariates

The possible covariates for the phenology analysis are based on climate data consisting of temperature, precipitation, elevation and geographical location. Time and geographical location also provides information on day length. For the DBB data the covariate accumulated growing degree days (agdd, defined in appendix A) is the most important and a variation of agdd; accumulated decreasing degree days (addd), for the leaf senescence.

The method of analysis in this thesis suggest that time direction in the covariates is of importance and therefore accumulated versions of the basic climate covariates such as precipitation and day length are also investigated.

Together with the varying covariates, time constant covariates containing climate information for the previous year are included. The idea is to investigate the effects of climate conditions on current phenological observations.

In addition interactions between time varying and constant covariates are included. However, interactions between varying and varying as well as between constant and constant covariates are excluded. The main reason is to avoid ambiguity of time references due to the nature of the implemented stochastic regression model. Informal experimental tests supported the exclusion of interaction between time dependent covariates.

A list of covariates and their definitions is given in Appendix A – Phenological Covariates.

6.2 Data selection

The training data for the Finland analysis is chosen as 2/3 of the available phenology stations. It is noticeable that the remaining 1/3 validation data contains less than 100 observations (table 2.1) making the prediction intervals sensitive to single observations.

For the UK a smaller subset 1/3 (DBB) and 1/4 (LS) is used due to computer memory restrictions. The smaller ratio for LS despite less total observations than DBB is due to the way the observations becomes larger due to later date of occurrence and since all sequences encoding start at January 1.

6.3 Parameter estimates

The models are estimated using the elastic net regularization with $\alpha \in \{0.6, 0.8, 1.0\}$ (chapter 4.3). The regularization shrinks the coefficients and give a sequence of models with decreasing model sizes (not strictly decreasing) as seen in figure 6.3a-b.Furthermore the obtained model sequences are re-estimated using the relaxed Lasso (chapter 4.4) for $\phi = 0$ corresponding to a GLM estimation of the model. The obtained model sequences from the regularization are cross validated with 10 groups in order to find the model with minimal CV error (min). Among the models within one standard error from the optimal CV-model the model having the fewest parameters is chosen as the optimal one standard error (1se) model. This choice is made among all elastic net models, CV-errors together with confidence bands are shown in figures 6.3c-f.

This gives, for the two countries (Finland and the United Kingdom), two phenological transitions (day of bud burst and leaf senescence), Lasso and relaxed Lasso, and min/1se model a total of $2^4 = 16$ possible models.

The validation results (table 6.4a-b) show that the one standard error relaxed Lasso models having smallest model sizes are as good as the optimal CV elastic net models. The only exception is the Finland LS model. In table 6.3a the covariates for the smallest models are shown.

A type of consistency test of the various regularizing methods is to see if the smaller models as generated by varying the λ parameter contain the covariates of the larger models.

Intuitively one would like the smaller models to have all their covariates in common with the larger models.

In general the covariates in the smallest models are also in the larger models. Table 6.3b lists the missing covariates for the different models. Most of the missing covariates are interactions with agdd (DBB) or addd (LS) which have been replaced by similar covariates. Overall the models seem to be consistent.

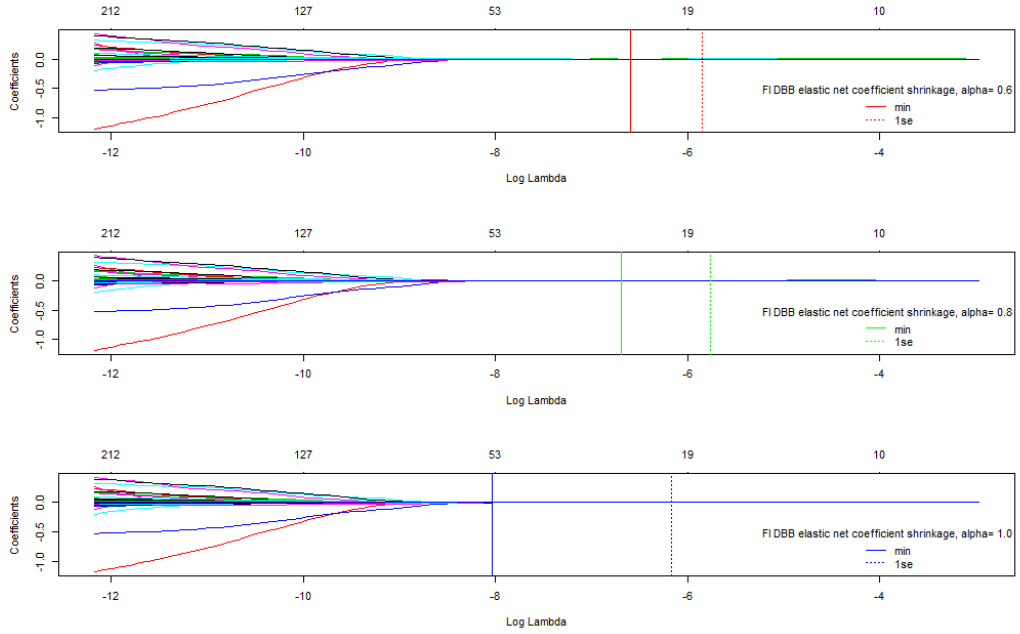


Figure 6.3a Finland DBB coefficient shrinkage for the different elastic net models. It's clear that smaller values of alpha keep coefficients longer.

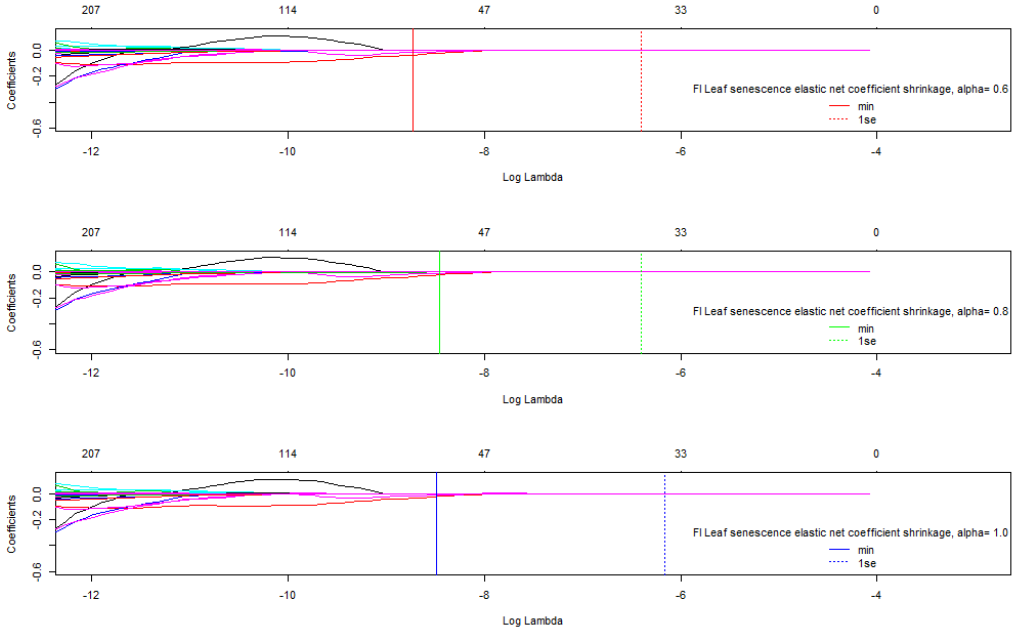


Figure 6.3b Finland leaf senescence coefficient shrinkage for the different elastic net models. It's clear that smaller values of alpha keep coefficients longer. The coefficient paths are more irregular than for the DBB.

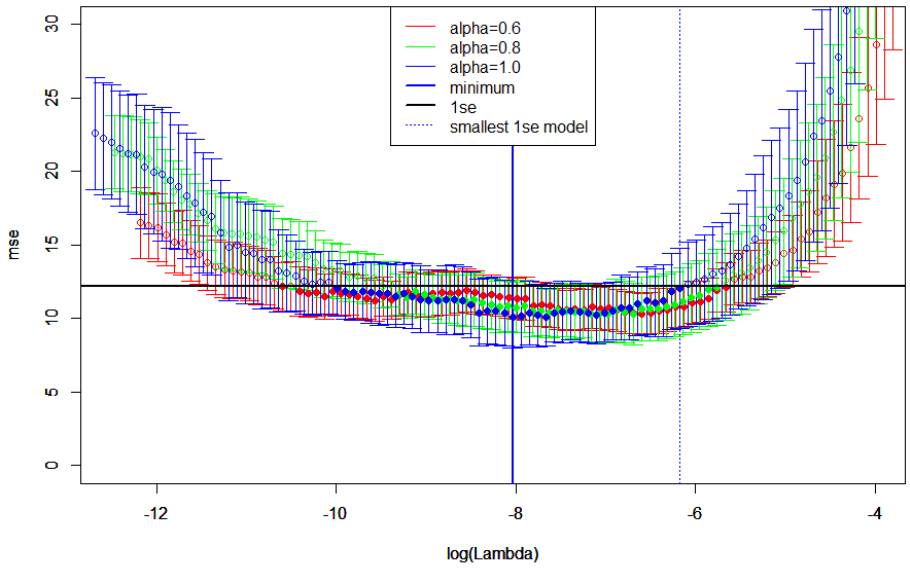


Figure 6.3c Finland DBB cross validation plots for regularization penalty parameter lambda and the standard errors. Solid colored circles are those lambdas within one standard error from the minimum. The smallest error is for elastic net parameter alpha=1.0. The smallest models for the three choices of alpha (0.6, 0.8, 1.0) is respectively (18, 17, 12). Consequently the smallest model within 1se is for alpha=1.0 (Lasso). It's clear that the elastic net keeps more and more covariates as alpha decreases.

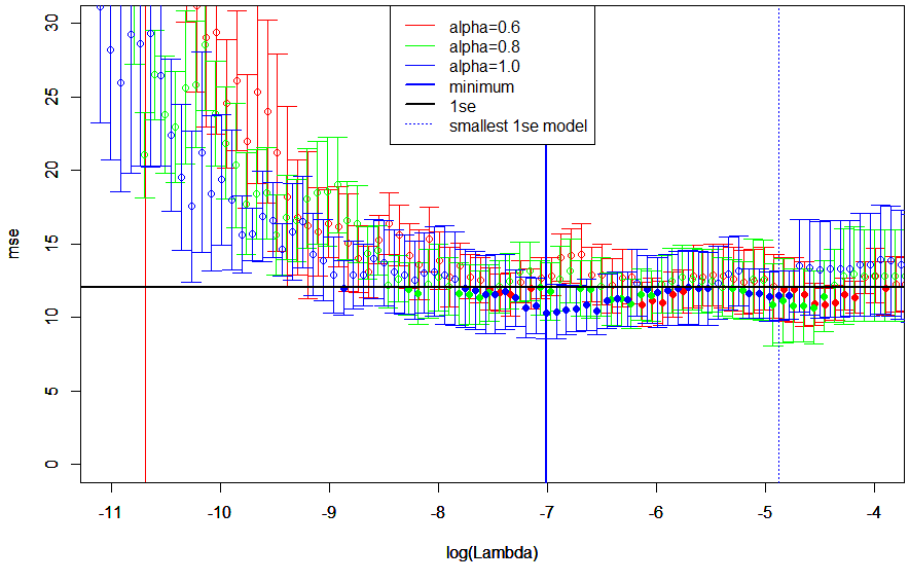


Figure 6.3d Finland DBB cross validation plots for regularization penalty parameter lambda and the standard errors for the ***relaxed Lasso*** models (re-estimated models). Solid colored circles are those lambdas within one standard error from the minimum. The smallest error is for elastic net parameter alpha=1.0. The smallest models for the three choices of alpha (0.6, 0.8, 1.0) is respectively (11, 10, 6). Consequently the smallest model within .1se is for alpha=1.0 (Lasso). It's clear that the relaxed Lasso reduces the model sizes.

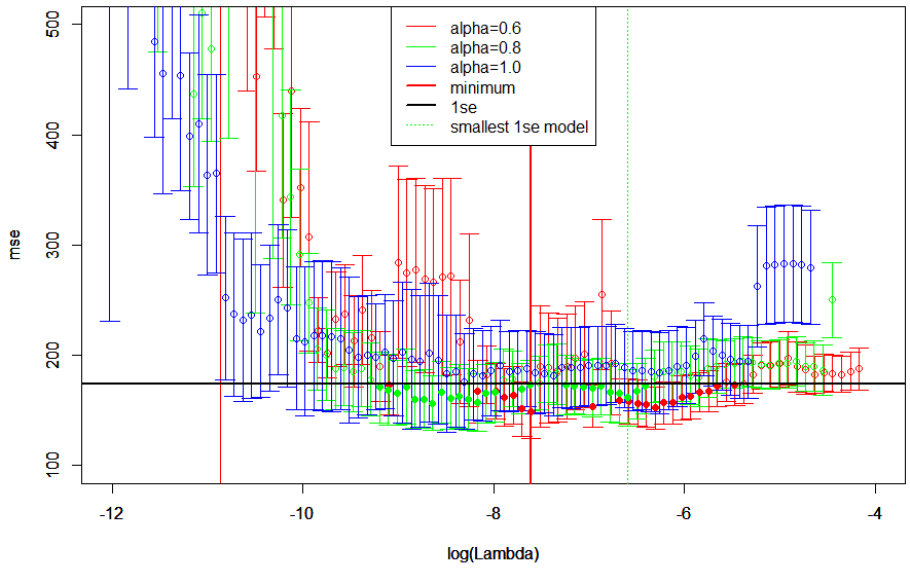


Figure 6.3e Finland leaf senescence cross validation plots for regularization penalty parameter lambda and the standard errors for the ***relaxed Lasso*** models (re-estimated models). Solid colored circles are those lambdas within one standard error from the minimum. The smallest error is for elastic net parameter alpha=0.8. Note that no model for alpha=1.0 is within one standard error. The smallest models for the three choices of alpha (0.6, 0.8, 1.0) is respectively (31, 29, NA). Consequently the smallest model within 1se is for alpha=0.8. It's clear that the relaxed Lasso reduces the model sizes, but not always within one standard error.

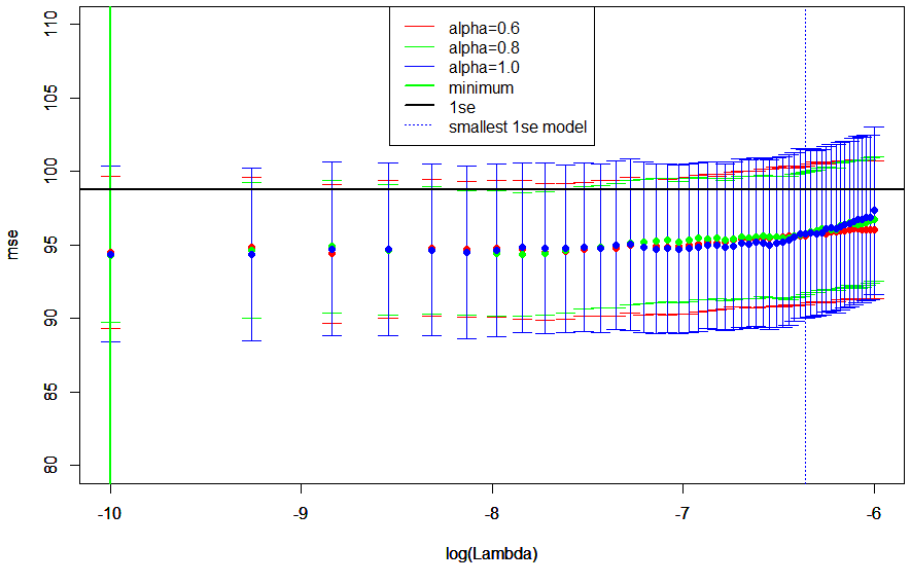


Figure 6.3f UK DBB cross validation plots for regularization penalty parameter lambda and the standard errors. Solid colored circles are those lambdas within one standard error from the minimum. The smallest error is for elastic net parameter alpha=0.8 but clearly the difference from the other alpha parameters is very small. In fact, due to the large standard errors for all three models, all choices of the lambda are within one standard error. The smallest models for the three choices of alpha (0.6, 0.8, 1.0) is respectively (32, 20, 8). Consequently the smallest model within .1se is for alpha=1.0 (Lasso). It's clear that the elastic net keeps more and more covariates as alpha decreases.

Table 6.3a The covariates for the smallest models.

Finland DBB	United Kingdom DBB	Finland LS	United Kingdom LS
agdd5 continentality.cic:agdd5 chill10:agdd5 number.spring:agdd5 mean.year:growthseason	adaylength chill5:adaylength continentality.cic:agdd5 chill5:agdd5 chill10:agdd5 number.spring:agdd5	addd26 addd28 chillm5:agdd5 mean.year:addd22 continentality.cic:addd26 total.spring:addd26 continentality.cic:addd28 total.spring:addd28 number.summer:addd28 number.fall:addd28	daylength:number.fall chillm5:agdd5 continentality.cic:addd22 continentality.cig:addd22 chill0:addd22 mean.year:addd22 mean.fall:addd22 chill5:addd26 mean.fall:addd26 number.spring:addd26 mean.fall:addd28 number.spring:addd28 total.summer:lastfrost

Table 6.3b Covariates in the smallest models for the Finland and UK DBB and senescence models not contained in the larger models.

	FI DBB	*FI LS*	*UK DBB*	*UK LS*
Lasso min	chill10:agdd5	number.summer:addd28 number.fall:addd28 chillm5:agdd5	adaylength chill10:agdd5	continentality.cic:addd22 continentality.cig:addd22 mean.year:addd22
Lasso 1se	agdd5	**Smallest model**	chill10:agdd5	continentality.cic:addd22 number.spring:addd26
Reest min	agdd5	number.summer:addd28 number.fall:addd28		
Reest 1se	**Smallest model**		**Smallest model**	**Smallest model**

6.4 Predictions

The estimated models in section 6.3 are evaluated on the validation sets and the result is shown in tables 6.4a-b. The prediction interval coverage is good (figure 6.4d), except for the Finland leaf senescence model. The Finland leaf senescence model is actually the only model having better R2 in the validation set, R2=0.49, than in the training set, R2=0.30 clearly indicating a mismatch between the validation and training sets. The low R2 also suggests a generally bad fit.

For all models, except FI LS, the optimal alpha value in the elastic net for the one standard error approximated models is 1.0 corresponding to a standard Lasso. This indicates that there is no need for elastic net regularization if one only seeks to find the most compact 1se model. The model sizes are larger for the Lasso than the relaxed Lasso models (except for Finland leaf senescence) supporting the conclusion of the relaxed Lasso (chapter 4.4). Note that the models are chosen using CV with weighting of the RMSE for the 10 different CV-models. The choice of optimal model and the one standard error imply that the RMSE for the optimal model is always smaller than the one standard error model, but since the model selection is made on the training set and the final validation set is different this is not always the case (tables 6.4a-b).

As seen in figure 6.4a the predictions for the Finland DBB follows the observation curve very closely, with R2 above 0.9 giving high confidence for the validity of the model. The prediction for the Finland leaf senescence, figure 6.4b, on the other hand fails to follow the observation curve at the beginning and the end. This suggests that the fit is mainly approximately averaged over all observations.

Not surprisingly all the models for the UK data sets fail to follow the observation curve due to the large variability in the data sets for observations within each climate grid cell, see figures 2.1a-b. The climate data cannot differentiate between different observations in a satisfying way. As a consequence all predictions look similar to the one in figure 6.4c.

Table 6.4a Prediction results on the ***validation data*** for the Lasso optimal (min) and one standard error approximated (1se) as well as for the corresponding relaxed Lasso models.

	MAE	RMSE	Bias	95 % PI coverage	95 % PI length	R2	R2 Max	Model size	α
Finland DBB									
Null Model	10.4	12.0	4.4	100%	46.1	-0.15		1	
Lasso .min	2.6	3.4	0.9	96%	12.5	0.90	1.0	29	1.0
Lasso .1se	2.5	3.3	0.8	96%	13.6	0.91	1.0	12	1.0
Relaxed Lasso .min	2.7	3.5	1.0	95%	12.6	0.90	1.0	16	1.0
Relaxed Lasso .1se	2.6	3.3	1.2	97%	13.3	0.91	1.0	6	1.0
UK DBB									
Null Model	9.3	11.8	0.0	95%	44.9	0.00		1	
Lasso .min	7.8	10.2	0.7	94%	38.1	0.25	0.51	82	0.8
Lasso .1se	7.9	10.3	-0.5	94%	38.4	0.23	0.51	8	1.0
Relaxed Lasso .min	7.9	10.4	-0.1	93%	38.0	0.22	0.51	38	0.6
Relaxed Lasso .1se	8.0	10.4	1.1	94%	38.3	0.22	0.51	7	1.0
FI Leaf senescence									
Null Model	10.2	12.3	2.2	99%	57.3	0.00		1	
Lasso .min	7.2	8.6	2.1	100%	50.7	0.49	1.0	47	0.8
Lasso .1se	7.5	9.3	-0.4	100%	54.1	0.42	1.0	11	1.0
Relaxed Lasso .min	8.2	10.4	2.6	98%	47.9	0.26	1.0	49	0.6
Relaxed Lasso .1se	7.4	9.3	1.1	100%	49.8	0.42	1.0	29	0.8
UK Leaf senescence									
Null Model	12.3	15.8	-0.2	95%	63.5	0.00		1	
Lasso .min	11.3	14.7	1.1	95%	59.2	0.13	0.48	67	0.6
Lasso .1se	11.6	15.0	0.5	95%	60.3	0.10	0.48	16	1.0
Relaxed Lasso .min	11.7	15.1	0.8	94%	58.9	0.09	0.48	30	0.8
Relaxed Lasso .1se	11.6	14.9	1.3	95%	59.9	0.11	0.48	14	1.0

Table 6.4b Prediction results on the ***regression data*** (data that estimated the coefficients)

	MAE	RMSE	Bias	95 % PI coverage	95 % PI length	R2	R2 Max	Model size	α
Finland DBB									
Null Model	9.4	11.8	0.4	96%	46.1	0		1	
Lasso .min	2.2	2.9	-0.4	96%	12.5	0.94	1.0	29	1.0
Lasso .1se	2.5	3.3	-0.7	95%	13.6	0.92	1.0	12	1.0
Relaxed Lasso .min	2.3	3.0	-0.3	96%	12.6	0.94	1.0	16	1.0
Relaxed Lasso .1se	2.4	3.3	-0.3	93%	13.3	0.92	1.0	6	1.0
UK DBB									
Null Model	8.9	11.5	-0.4	95%	44.9	0		1	
Lasso .min	7.3	9.5	0.6	96%	38.1	0.31	0.55	82	0.8
Lasso .1se	7.5	9.8	-0.7	95%	38.4	0.27	0.55	8	1.0
Relaxed Lasso .min	7.4	9.7	0.1	94%	38.0	0.29	0.55	38	0.6
Relaxed Lasso .1se	7.5	9.8	0.9	95%	38.3	0.27	0.55	7	1.0
FI Leaf senescence									
Null Model	12.7	14.6	0.4	97%	57.3	0		1	
Lasso .min	9.6	12.2	1.0	98%	50.7	0.30	1.0	47	0.8
Lasso .1se	10.9	13.3	-1.3	98%	54.1	0.17	1.0	11	1.0
Relaxed Lasso .min	8.0	10.6	1.3	98%	47.9	0.47	1.0	49	0.6
Relaxed Lasso .1se	8.6	11.5	1.4	97%	49.8	0.38	1.0	29	0.8
UK Leaf senescence									
Null Model	12.5	16.2	-0.4	95%	63.5	0		1	
Lasso .min	11.4	14.9	1.5	95%	59.2	0.15	0.51	67	0.6
Lasso .1se	11.6	15.2	0.9	96%	60.3	0.11	0.51	16	1.0
Relaxed Lasso .min	11.4	14.8	1.5	95%	58.9	0.16	0.51	30	0.8
Relaxed Lasso .1se	11.5	15.1	1.9	95%	59.9	0.13	0.51	14	1.0

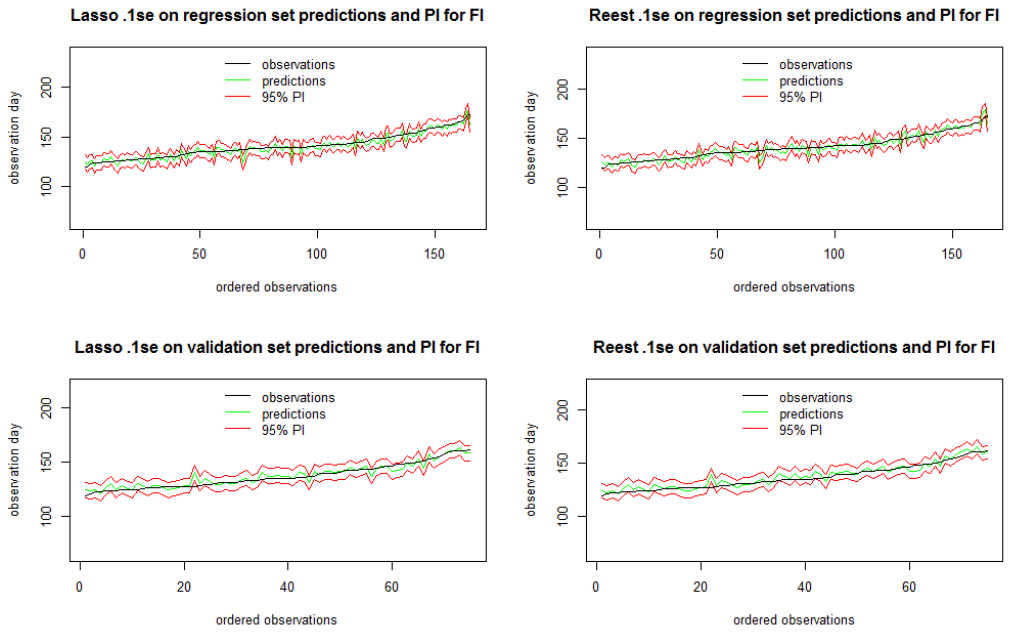


Figure 6.4a DBB prediction for both ***validation*** and ***regression*** set for Finland .1se models with original observations as the black line (ordered from earliest to latest). The predictions follow the original observations very well, with R2 0.91-0.92.

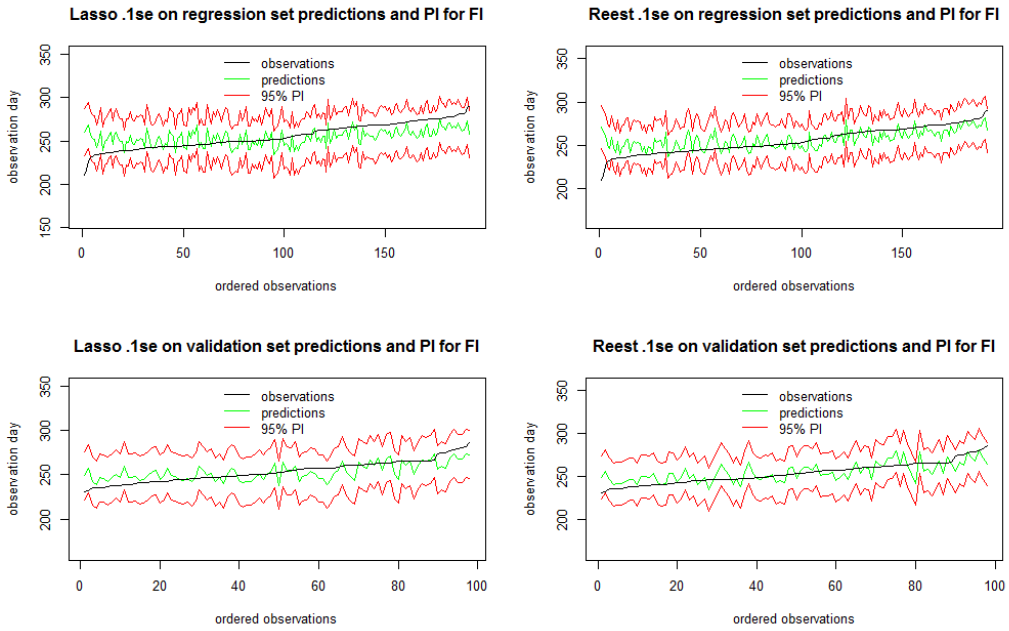


Figure 6.4b Leaf senescence prediction for both ***validation*** and ***regression*** set for Finland .1se models with original observations as the black line (ordered from earliest to latest). The predictions follow the original observations somewhat in the center but fails toward the endpoints. The PI is larger than for DBB and the models differ much more, with R2 varying from 0.17-0.42.

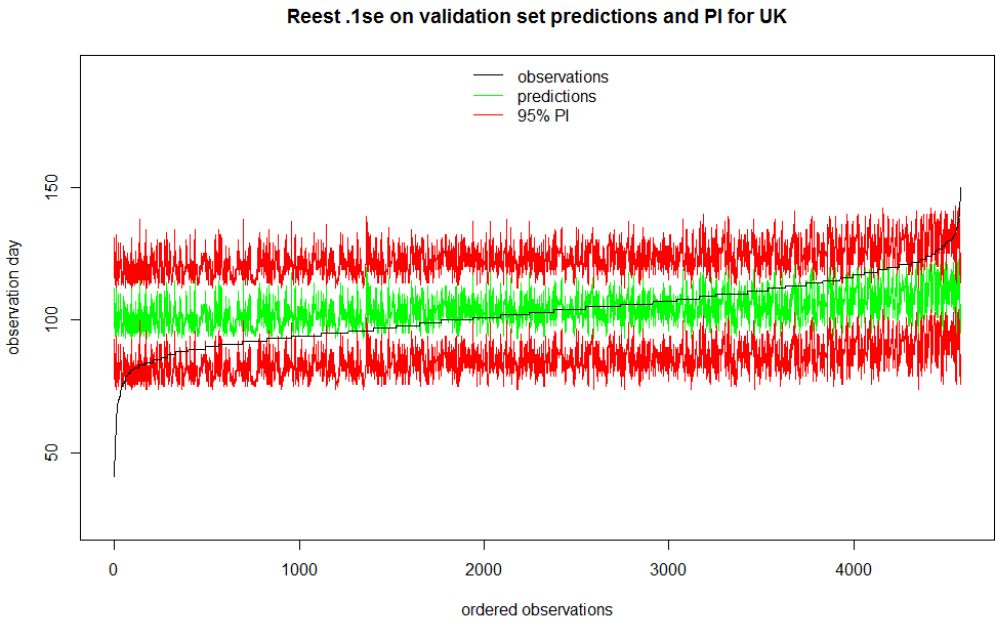


Figure 6.4c DBB prediction for the ***validation*** set for the UK .1se model with original observations as the black line (ordered from earliest to latest). Clearly the predictions do not follow the black line and the R2 is only slightly above 0.2. The prediction curves for all other UK models are similar.

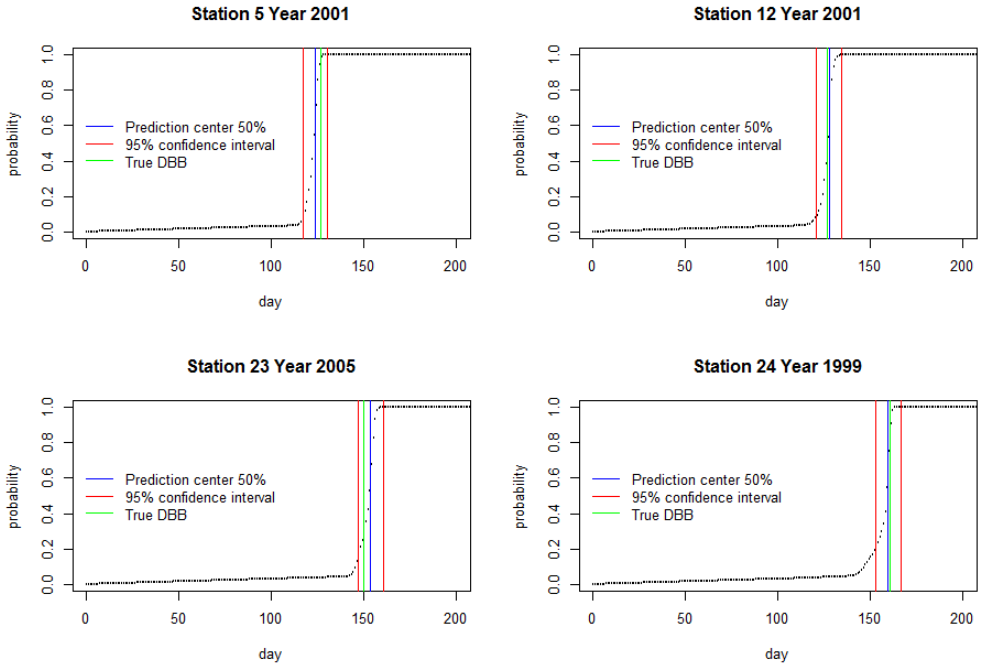


Figure 6.4d Prediction plots for the Finland DBB. The prediction slope is steep and PI narrow.

6.5 Attempts to improve the model

For the leaf senescence data in Finland several test where made to find better covariates. One was to start each sequence from the time of DBB (using the subset of observations that coincided), but it seems that the alignment of the time dependent covariates failed in the model (producing models not better than the null model). However test showed that different starting days for the DBB observations, by removing 75% of the days leading up to the DBB, gave as good a result as the analysis on full data (Song's model works but needs good covariates which growing degree days is for DBB). The implemented version of a constant threshold date at day 150, for the covariate addd, together with observation start at 1 January gave a result as good as any other. The addd is calculated from day 150 and set as zero for days 1-149. Different covariates were also constructed to try to find better models. Accumulated rainy days as well as non-rainy days and accumulated day length of non-rainy days etc. were tried. It seemed like the version of accumulated decreasing sunny days (or accumulated increasing non-sunny time) had statistical significance, as well as some other variants. The idea was that precipitation could give some information on cloudiness.

However none of the models gave better predictions than those presented here and they are therefore left out of this presentation. More direct information on actual sun exposure could provide a statistically relevant covariate, but then estimations of cloudiness would be required for future predictions.

Summary

The proposed models for regularization provide consistent covariate selections in the sense that all the larger models contain most of the covariates in the minimal model. The predictions remain as good for the smallest models as in the optimal CV models and the smallest model can therefore be used for predictions. In this case the smallest model corresponded to the one standard error relaxed Lasso model. The growing degree day covariate proved to be a very good explanatory variable for DBB. The model(s) for leaf senescence on the other hand had R2 below 0.5. Examining the prediction plots and results clearly showed lack of fit and the use of addd cannot conclusively be attributed as an influencing factor of leaf senescence. For the UK, having a large variability within each grid, the analysis could clearly not do much. R2 of the UK models were below 0.3 and in parity with some of the FI leaf senescence models, having R2 between 0.17 and 0.49. See figures 6.4b-c.

7. Analysis – Examination of engineers

For the modeling of the engineer's path to exam it is clear that if the covariates have exact information on how many points have been taken and when, the prediction of examination would be trivial and the model would be perfect. Nonetheless such a model could be efficient in determining factors affecting time to examination. Some might hasten (completing many points the first year) or prolong (completing few points the first year) the time to exam.

The analysis of the examination data will focus on using the available information to build a model that can capture the variation as well as possible. The part of the data concerning the examination has much of the variation in the amount of points taken since there is a required amount of points that each student needs to have. The information on student inactivity is modeled directly through constant covariates. For the graduating students the variation on inactivity captures random information not really easily model able otherwise. For the quitting students inactivity is an integral part since it makes up the most of the registered studying time, and so providing the exact number of inactive semesters more or less tells the regression when the student was de-registered (quitted). Hence the model will disregard most of the other data in the analysis making classification between graduated and quitted students impossible. Therefore the model for quit does not use the full information on inactivity more than for the dual model analyzed for predictions, instead a weaker indicator version is implemented. The constant covariate for inactivity is in the indicator case either 0 or 1 standing for has not had inactivity and has had inactivity respectively.

7.1 Data preprocessing and covariate selection

For the students who graduate the registration of the exam is often delayed to a semester after they finished. This delay can be several years. Since our main interest is when a student is formally eligible to graduate the examination is moved back in time to the last semester during which points were taken.

Another issue is the points for foreign studies. Examining the data and assuming (for students starting before 2007) a program points total of 270hp it seems reasonable to assign 30hp for foreign studies when no points have been registered. However, if points have been registered no adjustment is made. The correction for the exchange students is made before the correction of the time of exam. Finally cases of students quitting during the first semester are removed, since there is no information to base predictions on in those cases. The cases of students having a previous exam with points and therefore graduating with total accumulated points less than 270hp are kept and constitute one source of uncertainty in the model.

To fit the data to Song's model all students are assumed to start at the same time lined up beside each other. This is much the same as for the tree analysis where each observation year is treated similarly. The covariates are placed in three groups as varying, constant or categorical. Interactions are taken between varying and constant covariates much the same as

for the phenology analysis and finally interactions between categorical and both the constant and varying covariates are formed. For the examination data there are three constant covariates measuring the number of study breaks, foreign study semesters and inactive semesters. The different covariates used in the analysis are listed in Appendix B.

7.2 Program grouping and categorical standardization

Since there are 14 different programs the models generated with the programs as categorical covariates tends to be very large. It is therefore of interest to group related engineering programs. One way of achieving that is by applying GLM to the part of the data set where students have graduated. Since the program covariates are categorical one program will be left out. The idea is to take the coefficients of the program interactions and place them in a matrix. The left out program in the GLM regression will then have zero coefficients for its entire row. One then performs *k-means* (Hartigan, 1975) clustering of the programs according to their coefficients. Similar programs will hopefully have similar coefficients. A new model is then fitted using the clustered programs and one can assess the predictive quality of the new smaller model. In this way the model for examination was reduced from about 80 to 40 covariates.

A consideration when using glmnet regularization is the standardization of the covariates. In order for the regularizations to work optimally each covariate needs to have the same scale. However, categorical variables should not be scaled in order to preserve coefficient interpretability. If one scales categorical variables then as in this case, programs will have coefficients at different scale, implying Lasso penalties that depend on the number of students in each program.

It is also possible to have full representation of categorical covariates if one removes the regression intercept. This is done in some cases and when using the full set of categorical covariates it is possible to implement at the same time another group of categorical covariates by leaving one of the categories out as done for the female/male categorical covariates.

7.3 Models

The covariates for the engineer's data are presented in Appendix B. The constant covariate for study inactivity counts the number of inactive semesters for the student. This proves necessary for good modeling of graduating students. Intuitively inactivity is hard to model in any other way using the available data. For the quitting student model, on the other hand, that covariate explains too much and the predictions become almost deterministically exact. For classification the balance between graduation and quitting will be over biased towards quitting. Therefore in those cases, the covariate for inactivity is used as an indicator only. The models analyzed are:

Exam/Quit 1: Programs are not used as covariates, all other covariates are used. Since both male and female are used the model is without intercept. Total number of covariates is $3 \cdot 5 + 3 + 5 + 2 \cdot (3 + 5) = 39$.

Exam/Quit 2: Programs without clustering and without the male covariate. The intercept is included. The fit is made with only 1/20 of the data. The covariate for inactivity set as indicator. Total number of covariates is $1 + 3 \cdot 5 + 3 + 5 + (13 + 1) \cdot (3 + 5) = 136$.

Exam Group3: The program covariates in 3 groups and all other covariates except categorical male. The intercept is not included. Total number of covariates is $3 \cdot 5 + 3 + 5 + (3 + 1) \cdot (3 + 5) = 57$.

Exam All Programs: All (14) categorical program covariates and all other covariates except categorical male. The intercept is not included. Total number of covariates is $3 \cdot 5 + 3 + 5 + (14 + 1) \cdot (3 + 5) = 143$

Quit: Same as for the Exam model. The covariate for inactivity set as indicator.

7.4 Parameter estimates

For all models the parameters are calculated using elastic net parameter $\alpha \in \{0.6, 0.8, 1.0\}$. The best models are computed in the same way as for the phenology analysis. The second dual model is used for classification. Since the covariate for inactivity is set as an indicator the R2 for the exam part of the model becomes negative. This is because the model is too weak for predicting the examination time with good precision, and the model is not presented in the prediction tables. However for classification of students as graduating or quitting, the exact timing is of less importance.

CV standard errors, shown in figures 7.4a-d, are very small, and for the elastic net regularization the one standard error reduction of model size is negligible. However, the relaxed Lasso versions provide a substantial reduction but again the difference between minimal and one standard error model sizes is modest.

Covariates for the smallest model are presented in figures 7.5c-e. Note that even with grouping of programs the exam model has 38 covariates, compared to the full program model which has 90 covariates in the smallest model.

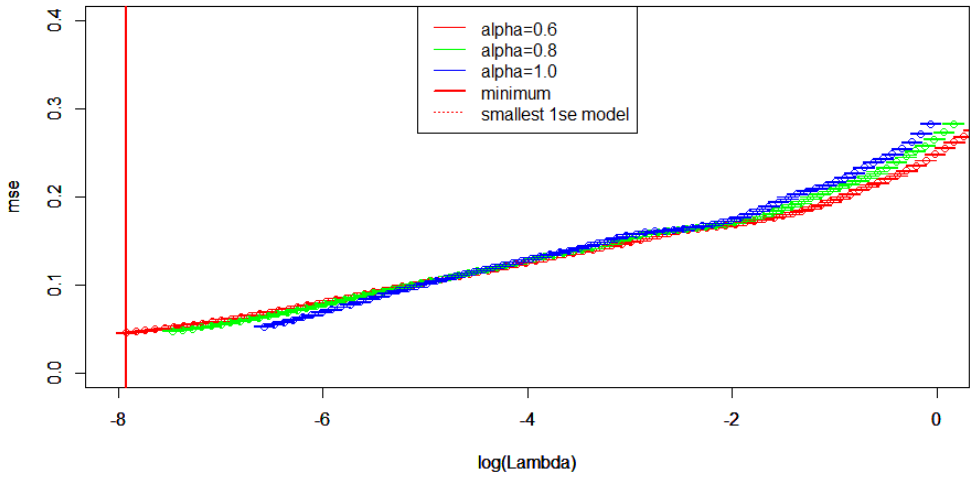


Figure 7.4a Regularization errors due to probabilities for the dual transition model exam/quit. The regularization favors a small penalty parameter λ and a small elastic net parameter α. The standard errors of the 10 fold CV is very small and the one standard error model is the optimal model itself.

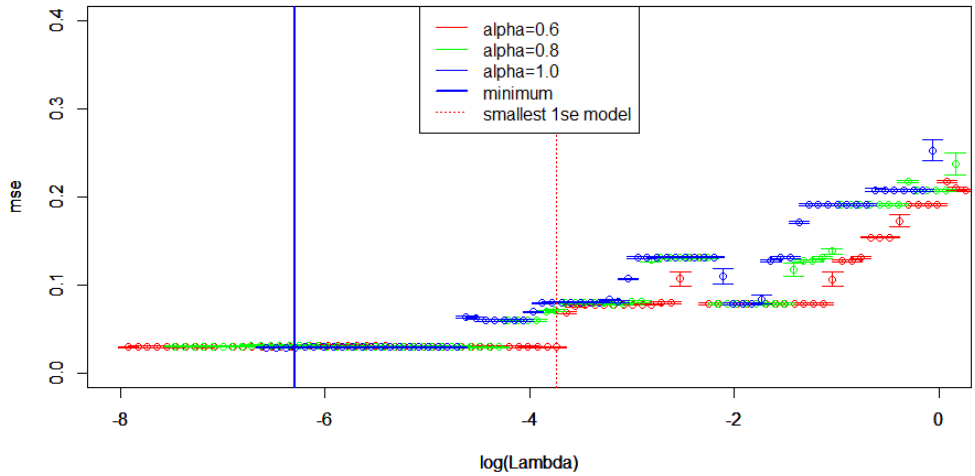


Figure 7.4b Regularization errors due to probabilities for the dual transition model exam/quit for the ***relaxed lasso***. The standard errors of the 10 fold CV are very small but since the errors are almost the same the one standard error model is different.

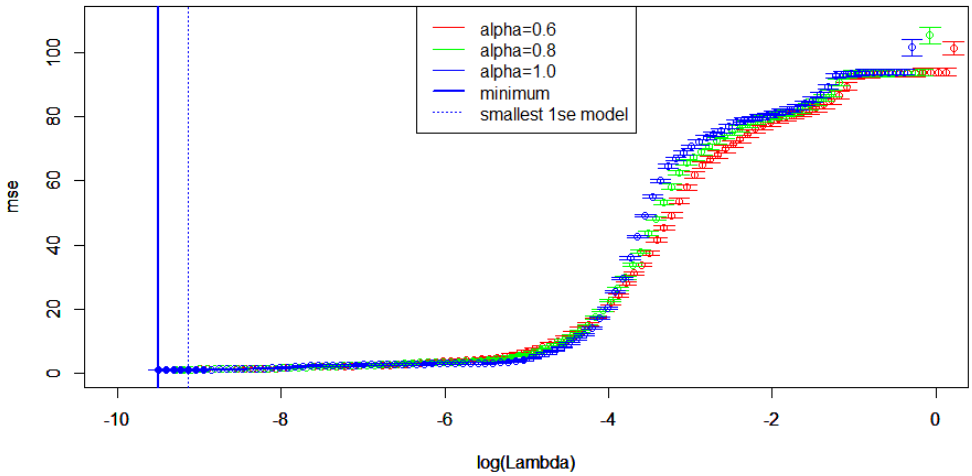


Figure 7.4c Regularization errors due to probabilities for the examination only model with programs in 3 groups. The regularization favors a small penalty parameter λ model.

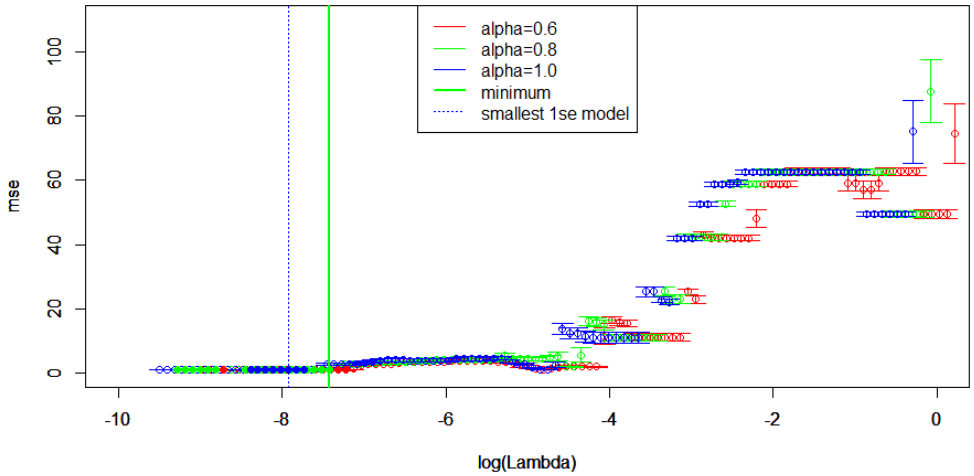


Figure 7.4d Regularization errors due to probabilities for the examination model for the ***relaxed lasso***.

7.5 Predictions & Classification

The prediction curve for the Quit model (figure 7.5a) looks similar to the one for the UK models in the phenology analysis (figure 6.4c). The corresponding prediction curve for the Exam model (figure 7.5b) is on the other hand similar, although not as smooth, as the prediction curve for the FI DBB analysis (figure 6.4a). The prediction results are shown in tables 7.5a-b. For the Exam and the Quit models the PI have a slightly too small prediction cover. The relaxed lasso provides the most compact models and all of the four model versions have comparable predictive power. The prediction for the Quit model has slightly worse results on the training set.

The predicted smallest models are shown in figures 7.5c-e. The model for examination has as many as 38 covariates. R2 for the full Exam model was between 0.85-0.87 slightly better than the Exam model with programs in 3 groups. Both of the Exam models used the same training and validation sets.

Classification is made using the two dual models. The prediction curves for the two transition models are compared at each semester and for each semester the curve with largest probability, or the one above the other, is used as classifier (figures 7.5f-g). This way even incomplete covariate sets such as students still studying can be classified. As seen in table 7.5c the classification only really works for the relaxed Lasso one standard error Exam/Quit 2 model. It gives good classification after semester 3 in the sense that for semester 3 and 4 any classification for quitting has large probability since the students graduating only have 4% misclassification. At semester 7-8, and after, the classification for graduation is good since the misclassification for the quitting students is low.

Table 7.5a Prediction result on ***validation*** set for the dual transition model and the single models.

	MAE	RMSE	Bias	95 % PI coverage	95 % PI length	R2	Model size	α
Exam/Quit 1(Quit)								
Null Model	8.5	10.4	-0.6	93%	40.5	0.00	1	
Lasso .min	0.1	0.7	0.1	99%	2.6	1.00	38	0.6
Lasso .1se	0.1	0.7	0.1	99%	2.6	1.00	38	0.6
Relaxed Lasso .min	0.03	0.3	0.0	98%	0.8	1.00	21	1.0
Relaxed Lasso .1se	0.01	0.1	0.0	99%	0.5	1.00	15	0.6
Exam/Quit 1(Exam)								
Null Model	1.6	2.5	-0.3	96%	10.4	-0.01	1	
Lasso .min	0.6	1.4	0.2	96%	5.5	0.68	38	0.6
Lasso .1se	0.6	1.4	0.2	96%	5.5	0.68	38	0.6
Relaxed Lasso .min	0.5	1.0	0.1	96%	5.2	0.83	21	1.0
Relaxed Lasso .1se	0.6	1.1	0.1	98%	4.5	0.80	15	0.6
Exam Group 3								
Null Model	1.7	2.6	-0.4	96%	10.2	-0.02	1	
Lasso .min	0.5	1.0	0.1	90%	4.0	0.83	45	1.0
Lasso .1se	0.5	1.1	0.1	96%	4.1	0.82	43	1.0
Relaxed Lasso .min	0.5	1.0	0.1	90%	3.5	0.86	39	0.8
Relaxed Lasso .1se	0.5	1.0	0.1	90%	3.6	0.86	38	1.0
Exam All Programs								
Null Model	1.7	2.6	-0.4	96%	10.2	-0.02	1	
Lasso .min	0.5	1.0	0.1	91%	3.9	0.85	106	1.0
Lasso .1se	0.5	1.0	0.1	91%	4.0	0.85	100	1.0
Relaxed Lasso .min	0.4	0.9	0.0	91%	3.5	0.87	93	1.0
Relaxed Lasso .1se	0.4	0.9	0.0	91%	3.6	0.87	90	1.0
Quit								
Null Model	8.4	10.3	-0.5	93%	40.7	0.00	1	
Lasso .min	3.6	6.1	-0.5	92%	20.2	0.64	21	0.8
Lasso .1se	3.8	6.3	-0.8	92%	20.5	0.63	8	1.0
Relaxed Lasso .min	3.6	6.1	0.1	93%	20.0	0.65	11	0.6
Relaxed Lasso .1se	3.7	6.2	0.0	92%	20.4	0.64	5	1.0

Table 7.5b Prediction result on ***regression*** or training set.

	MAE	RMSE	Bias	95 % PI coverage	95 % PI length	R2	Model size	α
Exam/Quit 1(Quit)								
Null Model	8.5	10.3	-0.4	93%	40.5	0.00	1	
Lasso .min	0.1	0.7	0.1	99%	2.6	1.00	38	0.6
Lasso .1se	0.1	0.7	0.1	99%	2.6	1.00	38	0.6
Relaxed Lasso .min	0.03	0.3	0.0	99%	0.8	1.00	21	1.0
Relaxed Lasso .1se	0.01	0.1	0.0	99%	0.5	1.00	15	0.6
Exam/Quit 1(Exam)								
Null Model	1.7	2.6	-0.4	96%	10.4	-0.02	1	
Lasso .min	0.6	1.4	0.2	96%	5.5	0.73	38	0.6
Lasso .1se	0.6	1.4	0.2	96%	5.5	0.73	38	0.6
Relaxed Lasso .min	0.5	1.0	0.1	97%	5.2	0.85	21	1.0
Relaxed Lasso .1se	0.6	1.1	0.1	97%	4.5	0.81	15	0.6
Exam Group 3								
Null Model	1.7	2.6	-0.3	96%	10.2	-0.02	1	
Lasso .min	0.5	1.0	0.0	91%	4.0	0.86	45	1.0
Lasso .1se	0.5	1.0	0.0	97%	4.1	0.85	43	1.0
Relaxed Lasso .min	0.4	0.9	0.0	92%	3.5	0.88	39	0.8
Relaxed Lasso .1se	0.4	0.9	0.0	92%	3.6	0.88	38	1.0
Exam All Programs								
Null Model	1.7	2.6	-0.3	96%	10.2	-0.02	1	
Lasso .min	0.4	1.0	0.0	92%	3.9	0.86	106	1.0
Lasso .1se	0.5	1.0	0.0	92%	4.0	0.86	100	1.0
Relaxed Lasso .min	0.4	0.9	0.0	92%	3.5	0.88	93	1.0
Relaxed Lasso .1se	0.4	0.9	0.0	92%	3.6	0.88	90	1.0
Quit								
Null Model	8.5	10.4	-0.5	93%	40.7	0.00	1	
Lasso .min	3.8	6.4	-0.7	91%	20.1	0.62	21	0.8
Lasso .1se	3.9	6.5	-1.0	92%	20.4	0.61	8	1.0
Relaxed Lasso .min	3.8	6.3	-0.1	92%	19.9	0.63	11	0.6
Relaxed Lasso .1se	3.8	6.4	-0.2	92%	20.3	0.62	5	1.0

Table 7.5c Classification results, using the two dual models. The classification percentage is calculated as the ratio of correct classification for both the graduation validation data and the quitting student's data. The relaxed Lasso one standard error model for the second dual model provides good classification from semester 3-4 especially for quitting classification since only 4% of graduating students get misclassified.

	Exam/Quit 1 , alpha=1.0 Semester 1-8 and classification %									Exam/Quit 2, alpha=0.6 Semester 1-8 and classification %							
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*		*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Lasso 1se Exam	1	2	3	9	19	41	59	73		0	0	0	0	1	8	16	39
Lasso 1se Quit	72	68	63	58	55	52	51	51		100	100	100	100	100	99	99	98
Lasso min Exam	Same as for Lasso 1se									0	0	0	0	1	7	14	35
Lasso min Quit										100	100	100	100	100	99	99	98
Reest 1se Exam	**37**	**38**	**37**	**38**	**58**	**80**	**82**	**88**		**91**	**97**	**96**	**96**	**87**	**83**	**72**	**72**
Reest 1se Quit	**65**	**62**	**63**	**60**	**58**	**57**	**55**	**55**		**10**	**36**	**67**	**76**	**87**	**90**	**94**	**95**
Reest min Exam	Same as for Reest 1se									0	1	1	3	8	14	20	51
Reest min Quit										98	98	98	98	98	98	98	98

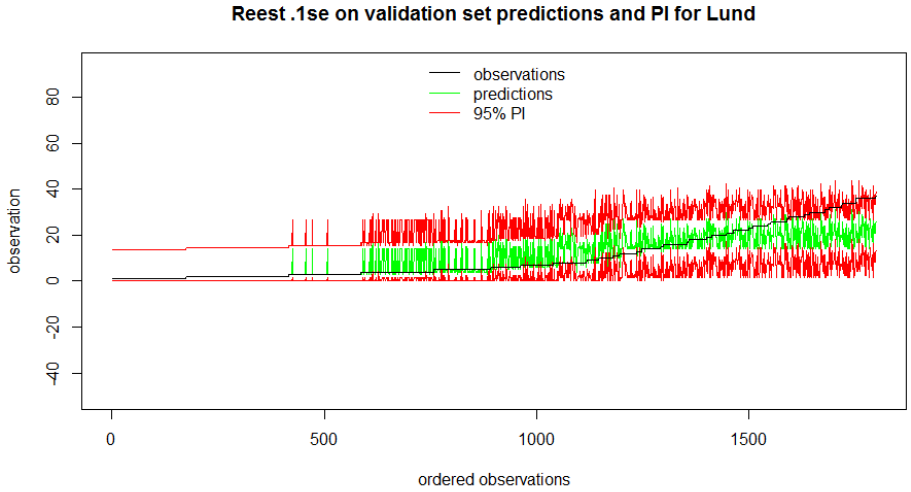


Figure 7.5a Prediction curve for the observations with PI for the quitting model. Clearly a R2 of 0.64 does not necessarily indicate a good fit.

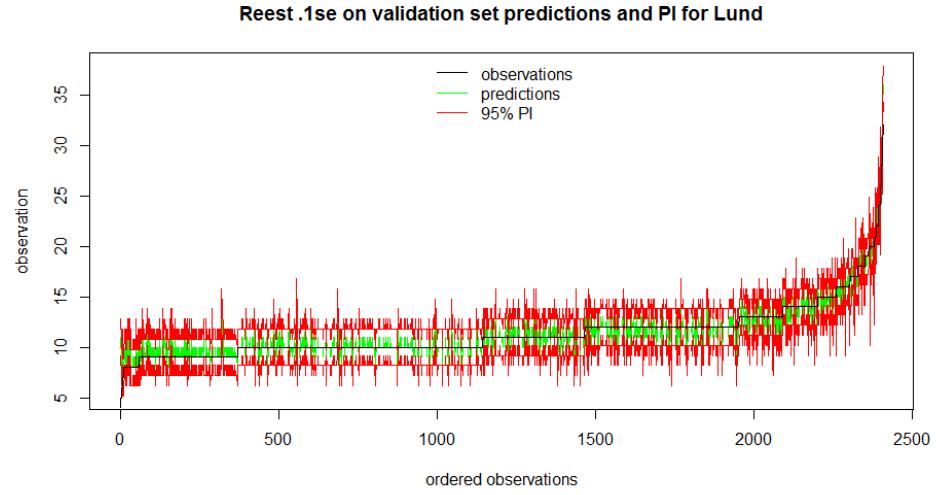


Figure 7.5b Prediction curve for the observations with PI for the examination model. The predictions follow the original observations fairly well and R2 is 0.85

	Exam	Quit
covariate	**coeff**	**coeff**
accp	6.61e-02	-1.39e-02
study.break	-1.69e+00	-1.36e+01
study.abroad	-6.95e-01	-5.39e-01
study.inactive	-1.49e+00	-1.32e+01
time	1.93e+00	1.31e+01
points	7.52e-02	-1.42e-02
regist	-3.06e-01	-3.28e+00
active	-1.33e+00	-1.28e+01
female	-2.35e+01	-5.37e+00
male	-2.11e+01	-5.55e+00
accp:study.inactive	-4.36e-04	4.57e-04
study.inactive:active	-4.07e-02	2.36e-04
accp:male	-3.09e-03	-1.02e-03
time:male	-2.01e-01	2.83e-02
study.inactive:male	2.49e-01	7.61e-04

Figure 7.5c The smallest dual transition model covariates. There is no intercept and the male/female covariates are the combined intercept. Since the model is dual the coefficients are not easily interpreted.

covariate	**coeff**	**covariate**	**coeff**
accp	6.28e-02	study.inactive:points	1.86e-02
study.break	-1.45e+00	study.break:regist	1.63e-02
study.abroad	-1.25e+00	study.abroad:regist	1.13e-01
study.inactive	-1.47e+00	study.inactive:regist	5.41e-02
time	2.08e+00	study.abroad:active	4.82e-01
points	5.97e-02	study.inactive:active	-6.47e-02
regist	-4.19e-01	accp:BGILMN	2.45e-02
active	-1.81e+00	accp:C	6.14e-02
female	-1.33e+00	time:female	1.57e-01
BGILMN	-2.51e+01	time:DEFKPVW	9.29e-02
C	-3.77e+01	points:BGILMN	-2.96e-03
DEFKPVW	-1.95e+01	points:C	1.40e-02
accp:study.abroad	-2.22e-02	regist:BGILMN	5.95e-02
accp:study.inactive	-1.78e-03	regist:C	-9.72e-02
study.break:time	-4.48e-02	active:C	4.05e-01
study.abroad:time	5.44e-02	study.break:DEFKPVW	-2.03e-01
study.inactive:time	1.64e-02	study.abroad:female	-2.12e-01
study.break:points	-1.45e-03	study.abroad:DEFKPVW	-3.20e-01
study.abroad:points	3.85e-02	study.inactive:DEFKPVW	-3.90e-01

Figure 7.5d Smallest Exam model.

covariate	**coeff**
study.break	-5.47e-01
study.inactive	-3.29e+00
time	5.55e-02
regist	-1.75e+01
study.break:time	1.77e-02

Figure 7.5e Smallest Quit model

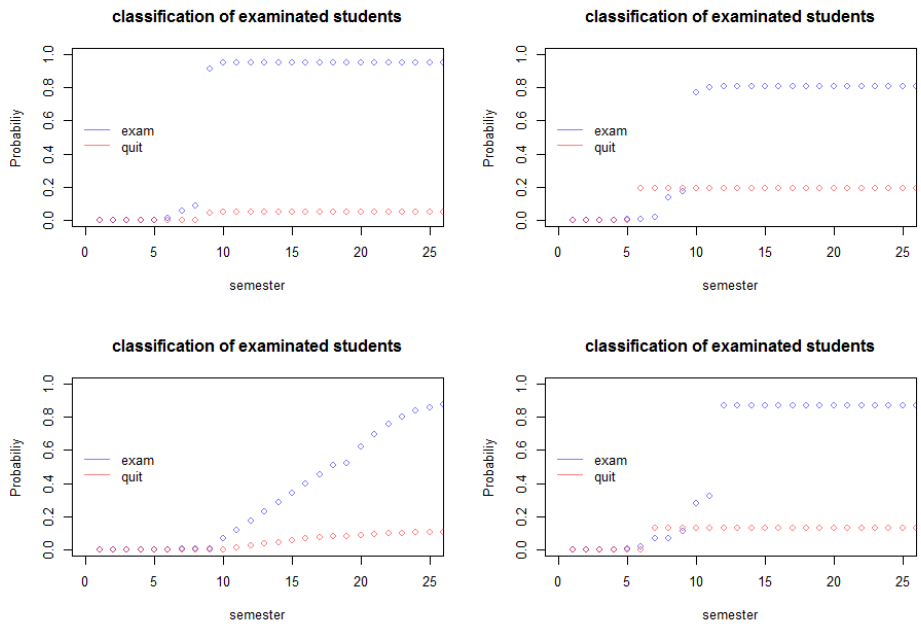


Figure 7.5f Classification curves for graduated students using the Exam/Quit 2 model for the one standard error relaxed Lasso. The examples are a random selection from the validation set. Misclassified quitted student at top right, between semesters 6 and 9, and bottom right, between semesters 7 and 9.

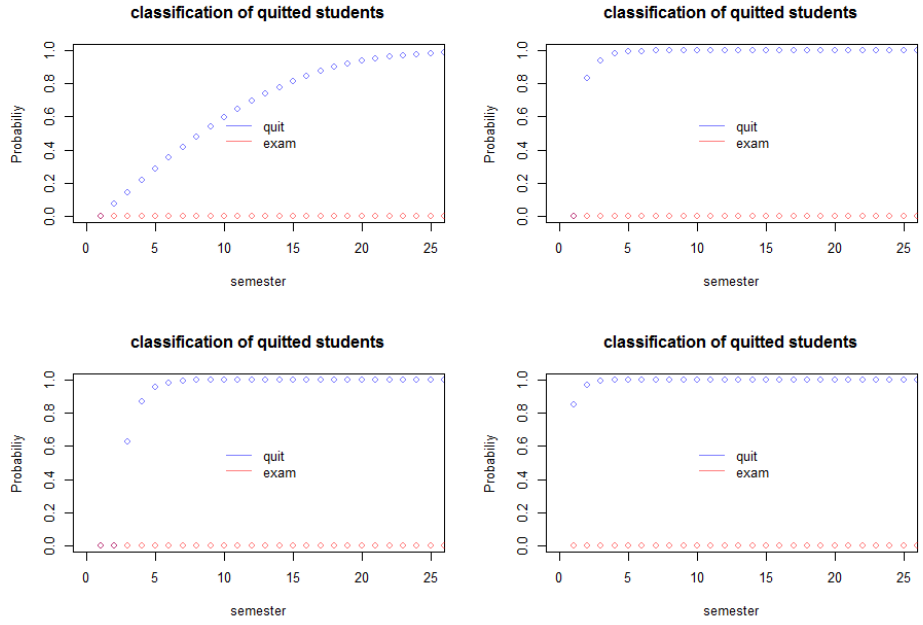


Figure 7.5g Classification curves for quitted students using the Exam/Quit 2 model for the one standard error relaxed Lasso. No clear cases of misclassifying of the graduated students. The examples are a random selection from the validation set.

7.6 Simulation

Among the students having graduated the ones with no inactivity, break or foreign studies and finishing exam within 10 semesters were selected. The students where then assigned one foreign study semester and predictions on exam time was made before and after. The difference in expected study time is presented in table 7.6. Surprisingly some students were actually expected to finish faster than without foreign studies. The Exam model for $\alpha = 1$ was used in the simulation.

Table 7.6 Simulation of expected time to exam for students without foreign studies. By assuming one foreign study term the predictions changed.

Unchanged	Slower	Faster	Total
2380	306	43	2729

Summary

The model of Song is transferrable to the examination data. The dual transition version allows for classification of the two absorbing states exam and quit. This enables a direct method of assessing the potential number of expected future examinations. The dual model with quantified inactive covariate proved to be too contaminated by future information to be a good classifier, a consequence of the model forcing the probabilities down for as many semesters defined as inactive. The weaker version on the other hand was efficient only when fit with a small number of training data contradicting the notion that more training data gives a better model. One explanation could be that the large amount of training data and unique covariate sets for each student possibly over fit the regularization process. Even so the standard errors will be extremely small (figures 7.4a-d) making the selection process, due to one standard error, ineffective. The classification could perhaps be improved by un-grouping the dual model covariates. Having the same covariates for both the examination and quitting transitions is probably not optimal. However test on the Lasso un-grouped dual regularization model showed no improvement in classification. The found classifier proved to be efficient and there were signs of improvement for smaller values of the elastic net parameter α. The model could perform simulations such as changing the status of students with no foreign studies to students with foreign studies and then compare the change in predicted examination time. On the other hand one could examine the classification as the foreign studies changes as well. It is easy to think of a variety of analyzing possibilities making the approach attractive.

8. Conclusions

The ‘Stochastic Process Based Regression Modeling of Time-to-event Data’ model of Song has been applied to both phenological data as well as to examination data. Song’s model could be efficiently extended to use many covariates as well as many transition states. The phenological analysis of Song was reproduced in this new setting and worked well. The sequential model of Song was not verified due to the lack of relevant covariate data. For the leaf senescence it was clear that the covariates could not provide an efficient model, probably due to lack of relevant climate data. The analysis of the Engineer’s path to examination or quitting was successfully modeled with Song’s model. The model for the quitting students was, as expected, not good for predictions due to the unpredictability of when the students quit. The model could capture the difference in a student about to quit or complete his/hers exam. The examination was well modeled even with incomplete data. The different regularization techniques proved to behave as expected and provides powerful tools in the aid of selecting suitable covariates.

8.1 Future work

The analysis of phenology can be explored deeper with Song’s model. The classification could perhaps be improved by implementing the un-grouped version (allowing the different transition states to have different covariates) of multinomial regularization using the relaxed Lasso. The prediction intervals need improvement, especially for discrete or coarse data. The examination model suggests that many phenomena in society could be modeled in the same way provided suitable covariates exist. The multiple transition models could perhaps be used for classification even for improper state transitions. One can imagine splitting the examination students into non overlapping groups taking into account the grading at examination. The probability of graduating with a specific grade could then be modeled although information on grading would be needed. The effect of having two similar covariates as points and grading is though not possible to foresee. Overall the ability to choose which covariates to use, what data to look at and in which way provides combinations of ever growing complexity making the work on any model extendable and improvable limited only by imagination and available data.

References

Akaike, H. (1974): "A New Look at the Statistical Model Identification," I.E.E.E. Transactions on Automatic Control, AC 19, 716-723.

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.

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

Delpierre N.; Dufrêne E.; Soudani K.; Ulrish E.; Cecchini S.; Boé J.; Francois C.; (2009) Modelling interannual and spatial variability of leaf senescence for three deciduous tree species in France. *agricultural and forest meteorology* 1 4 9 (2 0 0 9) 9 3 8 – 9 4 8

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/

Hart Hartigan igan J.A. (1975). Clustering algorithms

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

Meinshausen N. (2006) Relaxed Lasso. *Computational Statistics & Data Analysis 52 (2007) 374 – 393*

Montgomery, Douglas C. (2001). Design and Analysis of Experiments. New York: Wiley.

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 –2 ed.

R Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Schwarz, G. (1978): "Estimating the Dimension of a Model," Annals of Statistics, 6, 461-464.

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.

Tikhonov, A. N. (1943). "Об устойчивости обратных задач" [On the stability of inverse problems]. Doklady Akademii Nauk SSSR 39 (5): 195–198.

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

Zou H. Hastie T. (2004) Regularization and variable selection via the elastic net. *J. R. Statist. Soc. B (2005) 67, Part 2, pp. 301–320*

Appendix A– Phenological Covariates

Constant	Formula	Variable(s)
Chilling days last fall	Threshold values -5, 0, 5, 10 degrees Celsius. $T_t \in \{-5,0,5,10\},\ \text{chillT(n)} = \sum_{i=1}^{n} 1(T_{obs}(i) - T_t < 0)$ The number of days last fall having minimum temperature below thresholds. Calculated over Oct-Dec	**chillT**
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 $T_t \in \{-2,0,5\},\ \text{gddT(n)} = \sum_{i=1}^{n} \max(T_{obs}(i) - T_t, 0)$ degrees Celsius from January 1	**gddT**
Chilling degree days CDD	Accumulated cold temperature below a threshold $T_t \in \{-2,0,5\},\ \text{cddT(n)} = \sum_{i=1}^{n} \min(T_{obs}(i) - T_t, 0)$ in degrees Celsius from January 1	**cddT**
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**
Accumulated day length	Accumulated day length from 1 jan	**adaylength**
Accumulated rain	Accumulated amount of rain from 1 jan.	**arain**
Decreasing degree days	Accumulated temperature below a threshold 22, 26, 28 degree Celsius starting from day 150.	**adddT**

Appendix B– Student data Covariates

Constant	Formula	Variable(s)
Reported study break	Number of semesters student had a study break as an constant covariate	**study.break**
Reported abroad studies	Number of semesters the student had foreign studies as a constant covariate	**study.abroad**
Inactive	Either number of semesters student has been inactive or as an indicator for students having had at least one semester of inactivity	**study.inactive**
Varying	**Formula**	**Variable(s)**
Points taken	LADOK points for each semester	**points**
Accumulated points taken	$\text{accp(t)} = \sum_{i=1}^{t} \text{points(i)}$	**accp**
Time	The semester order from the beginning of studies 1,2,3…	**time**
Semester registration	$regist(t) \in \{1,...,10\}$ if no inactivity otherwise 0	**regist**
Activity	Accumulated active study time. All time except for reported quitting, study break and inactive semester with no study points registered.	**active**
Categorical	**Formula**	**Variable(s)**
Woman/man	Indicator if woman or man. Two different covariates.	**female, male**
Program	Indicator for program belonging. 14 different covariates.	**D E F K L M V I W B C P G N**
Grouped programs	Indicator for belonging to program group. Three different covariates.	**C BGILMN DEFKPVW**

Master's Theses in Mathematical Sciences 2013:E41
ISSN 1404-6342

LUNFMS-3048-2013

Mathematical Statistics
Centre for Mathematical Sciences
Lund University
Box 118, SE-221 00 Lund, Sweden

http://www.maths.lth.se/