Late Blight Prediction and Analysis

LUND UNIVERSITY
DEPARTMENT OF MATHEMATICAL STATISTICS

MASTER THESIS

Author:
A. Lukas Bengtsson

Supervisor:
Veiko Lehsten
Erik Lindström

June 27, 2017

Abstract

Phytophthora infestans cause severe losses in potato cultivation. Currently one of the most effective ways to reduce the related loss is using fungicides. For these to have optimal effect they must be applied approximately 10 days before the first visual sign of late blight. As the appearance of late blight is hard to predict there is heavy overcompensation of fungicides nowadays. By better forecasts less chemical compounds will be used lowering the environmental impact. It may also increase the profit from potato cultivation as less money goes to fungicides.

There are several models trying to predict the late blight appearance date of which one of the most well known is SIMCAST. Also Cox regression models have recently been looked into. Is it possible to improve these models? Moreover what financial impact will it have for farmers?

First an empirical investigation of the data was conducted for a general understanding of the subject. Then it was analyzed using, linear regression, Cox regression, Elastic net and cross-validation. To fill in missing data imputation was used.

The best performing model was obtained through Cox regression using δSIMCAST[1], PAL[2] and planting day as covariates. It had approximately the same failure ratio as foliage prediction which is currently used, but had 23% more accurate predictions and lower standard deviation. SIMCAST was principally out-performed by all other models. Using elastic net it was possible to obtain models which was approximately as good as the best model. Using the best model savings around 422 SEK/ha can be made from reduced fungicides alone[3].

The newly obtained models has potential to increase the profit of potato cultivation. Also by these the accuracy of prediction would contribute to

[1]This variable is defined within the thesis

[2]A measure used to estimate the need of phosphorus fertilization.

[3]This may be improved adjusting the safety threshold which was chosen such it was comparable to the currently used methods.

lower environmental impact. Notable is that the data used is not optimal for the experiment as some had to be estimated etc. hence before relying on these models field experiments should be considered.

Acknowledgements

What is a man without friends and family? I know that their support is vital to me thereby I'd like to thank them all. In particular I would like to thank Kevin, my faithful fellow mathematician with whom I have had a long journey through our studies and leaned me to work hard, steady and without loosing track of the goal.

This work would not be possible to do without the data which has been collected over time, for this and also a contagious burning spirit I want to give special thanks to Lars Wiik and co-workers. He also supplied photographs when needed and drowned me in additional articles of all kinds for which I am grateful.

Writing a thesis during long periods is not always easy as the path changes a long the way. Advanced well thought through ideas may prove to be total failures whereas the most simple way can the answer. Therefore I would like to thank all of my guides whom pointed me in the correct way when I was lost, in particular my supervisors Veiko Lehsten and Erik Lindstöm.

Contents

Page

1 Introduction **1**
1.1 Design . . . 2
1.2 Background . . . 2
1.2.1 Late blight . . . 3
1.2.2 Control . . . 4
1.2.3 Future Prognosis . . . 5
1.3 Bintje . . . 6
1.4 Limitations and Aim pf Thesis . . . 7
1.4.1 Questions at Issue . . . 7
1.4.2 Assumptions . . . 7

2 Method **9**
2.1 Multiple linear regression . . . 9
2.2 Cox regression . . . 10
2.3 SIMCAST . . . 11
2.4 Akaike's information criterion . . . 13
2.5 Imputation . . . 13
2.6 Elastic Net . . . 13
2.7 Cross-validation . . . 14

3 Results **15**
3.1 Data . . . 15
3.1.1 Potato cultivation data . . . 16
3.1.2 Weather data . . . 18
3.1.3 Data transformations . . . 18
3.2 Regression and illustrative plots . . . 20
3.3 SIMCAST . . . 24
3.4 Cox Regression . . . 27
3.5 Predictions . . . 30
3.5.1 SIMCAST, cSIMCAST and foliage prediction performance . . . 30
3.5.2 Cox prediction . . . 32
3.6 Elastic net cox regression . . . 34
3.7 Prognosis overview . . . 35
3.8 Profit using alternative models . . . 36
3.9 Late blight damage . . . 38
3.10 Linear regression . . . 38

4 Discussion **39**
4.1 Results . . . 39
4.1.1 Data . . . 39
4.1.2 Regression . . . 40
4.1.3 SIMCAST . . . 40
4.1.4 Cox Regression . . . 41
4.1.5 Predictions . . . 42
4.2 Profitability . . . 43
4.2.1 Late blight damage . . . 43
4.3 Question at Issue . . . 44
4.4 Data . . . 45
4.5 Perspective . . . 46
4.6 Continuations and improvements . . . 46
4.6.1 Mathematical considerations . . . 47

5 Summary **49**

Bibliography **51**

A SIMCAST **53**
A.1 SIMCAST . . . 53
A.2 cSIMCAST . . . 54
A.2.1 Beta SIMCAST fit . . . 54
A.2.2 Logistic SIMCAST fit . . . 54

Chapter 1

Introduction

The environmental strains are not rarely disputed in various manners. The well known greenhouse effect, which is featured by many researchers etc., is just one of many. The human kind spread all kinds of tracks behind, as they manipulate their environment focusing on optimizing profit. Some species are negatively affected whereas others gain unnatural advantages entailing abnormal populations [1].

A special case of such issue, which is not much spoken of in the everyday news, is the unnecessary usage of pesticides in potato cultivation . Many people are not aware of the extent nor possible effects of these. The prediction of pests are difficult, due to heavy dependence of factors such as weather and cultivation conditions, some may overcompensate the usage of fungicides. Notably is that generally from an economic point of view its preferable to have a safe strategy due to the loss increase from failure exceed those of excess spraying. However overcompensating implies unnecessary costs for the agriculture moreover has a negative impact to surroundings [2].

Late blight is one of the pests which has large economic impact, regarding potato cultivation, and will be the focus in this thesis [1]. Both controlling and refined breeding to improve resistance against late blight has proven to be difficult. However there has been positive results in late blight infection control which are used for decision making. The safety of using such programs is low due to lack of insurances against unsuccessful controlling attempts. This thesis aims to evaluate damage due to late blight and further evaluate forecasts of late blight.

1.1 Design

Initially, the subject will be introduced from a general angle of incidence based on related studies. Then the chosen methodology is defined and explained including Cox regression and its concepts, generalized linear models etc. By thoroughly investigating data, relations within and possibilities a foundation will be made for analysis. Then by using Cox regression to model appearance date of late blight using different treatment programs, generalized regression to find how it affects the yield and profit. After analysis a discussion follows and considers both the results and future possibilities to proceed towards defining a proper insurance. Lastly a summary and reference list will be given.

1.2 Background

There is a well known economic risk in cultivation of potato crops[3]. Late blight is no recently discovered phenomena and is probably one the causes behind disasters such as the great famine in Ireland during 1845-1847 [4][3]. As there is a high severity of late blight infections, many experiments and studies has been made in the subject [1]. This both regarding selective breeding to increase natural resistance, and control programs to delay and reduce the effect of an infection [2]. Moreover global environmental changes affects the species which cause late blight as well as the crops locally. Concerning Swedish experiments global warming seems to result in higher temperatures and more rain [5], which seem to have created a more favorable climate for some pests [1][6].

1.2.1 Late blight

Phytophthora infestans is a family oomocytes, fungus-like eukaryotic micro organisms, which mainly infects tomato and potato plants[2][3]. Its hosts suffer of damage on the foliage which known as late blight, and also on tubers called tuber blight[2]. Globally the loss in cultivation of potatoes is estimated to be 40%, of which 21% are due to diseases where a major factor is considered to be late blight[1].

There are several varieties of *Phytophthora infestans* which makes it hard to fully protect the crops [3]. It is very adaptable and rapidly create resistance against fungicides hence it is generally suggested to use a variety of treatments to avoid such scenario[2]. There are several kinds of treatments with distinct strengths and weaknesses used at different stages[6]. Typical constant covariates which are considered is the kind of soil, sandy soils indicates a lower infection rate as it does not bind the water whereas loamy are the opposite[6]. Optimal conditions are temperature around 15 - 20^{o}C and damp leaves during at least four hours, which may be frequently sampled [6][2]. Rain, wind and the growth rate of the potato tops are also relevant variables[6]. Notably is tops with a high growth rate tends to be a favorable micro-climate for the late blight[6].

Figure 1.1: Potato plants infected by late blight. The two bottom images are one resistant besides a susceptible species. Photo by *Lars Wiik*(2015-07-09)[7].

Figure (1.1) shows how plants are affected, also the large difference between species at a certain stage.

In order to minimize the loss due to this disease one focus is to delay the infestation day as long as possible [6]. Moreover it is suggested that spraying should be applied at least 7-14 days before the first late blight infection [6]. Naturally treatments differ in both cost and efficiency which may vary depending on weather conditions etc. [2].

1.2.2 Control

During 1983-2012 several experiments[1] were made aiming to control late blight infections by using different treatment programmes [2]. The natural resistance against *Phytophthora infestans* are considered to be low and the selective breeding has had lacking results[2]. These were some of the underlying reasons of the experiments and also why fungicides are commonly used to protect the harvest [2]. But as previously mentioned the *Phytophthora infestans* are healthy, aggressive and very adaptive, hence the issue is difficult[2]. During the conducted experiment occurrence date changed to an earlier time point [2].

Wiik's experimental results indicated that tuber blight was uncorrelated to the first appearance of the late blight, however there seemed to be some time dependence [2]. The experiments show that delaying the start of attack, and using fungicides increase usable harvest [2]. This is probably due to the foliage decay becomes slower, and thereby the plant may be productive for a longer duration [2].

The author of [2], *Lars Wiik*, emphasize that usage of fungicides might have negative environmental effects. He also claims that selective breeding is preferable, but due to lacking results in the area, timing and dosage are important in order to optimize the effect of fungicides whilst minimizing environmental impact. [2]. Regarding GMO, it might have a negative influence of the resistance against fungies as the genetic base will decrease according to *Wiik*[2]. The programs for spraying optimization which was tested proved to have a positive result hence should be considered a valid option until any better appear [2].

[1] This analysis is based on the same data hence more information about these will be given when introducing the data.

1.2.3 Future Prognosis

A global scale prognosis of late blight proliferation was made in *"Climate change may have limited effect on global risk of potato late blight"*[1]. They found that the global climate change may create a more favourable environment, for *Phytophthora infestans*, i.e. also the loss due to late blight will increase[1].

The prognosis was made using an amplified "*SIMCAST*"[2] model which takes the covariates of temperature and relative humidity in order to estimate the risk of damaging levels of late blight [1]. The prognosis suggested that there was a limited increasing loss due to late blight which is to be expected [1]. The authors emphasize that rather strong assumptions was made, such as stationary cultivation conditions.

[2]The basic SIMCAST will be more discussed further on.

1.3 Bintje

An old historically trusted potato sort is Bintje which is known to give large qualitative yield [3], however it is fairly susceptible to late blight. Due to the susceptibility and new options in sorts, the usage of Bintje has decreased [3]. Nevertheless not all replacing sorts is better with respect to resistance hence the issue caused by late blight is not assumed to disappear [3]. To summarize a few kinds[3]:

Bintje:

- High stable yield
- Medium early
- Good cooking properties
- Susceptible to late and tuber blight, 2 on a scale of five where one is the worst.

Bellona:

- High stable yield
- Bad cooking properties
- Medium susceptible to late and tuber blight, 3 on the scale

Elin:

- Medium yield
- Late
- Good cooking properties
- Medium susceptible to late and tuber blight, 3 on the scale

Note above is just a few samples of many kinds which has different properties. Fungicides and defoliation also affect the tuber quality and yield [3]. Despite the susceptibility of Bintje it is considered to be profitable [3].

[3]This information was obtained in [3].

1.4 Limitations and Aim pf Thesis

There is a large complexity in the stated issue due to dependence on several variables such as weather and cultivation conditions, hence the solution may not be unique nor trivial. The thesis is limited to the following questions at issue and assumptions:

1.4.1 Questions at Issue

- Would it be possible to achieve commercial gain by usage of the Cox model defined by Lehsten et al.?
- Regarding the model by Lehsten et al., is it possible to develop a better similar model?
- Comparing to SIMCAST, which method is preferable with respect to profit?
- How may the analysis be proceeded to improve results or defining a proper insurance?

1.4.2 Assumptions

The cultivation data is assumed to be representative, properly collected and the experiments well defined. Notably is that there lies a natural order in the data and its sampling frequency, example wise defoliation which is made slightly before the harvest, hence also is regarded in the analysis. Weather related data is assumed to be both accurate as measured data and 10-days ahead predicted data. Assumptions related to models will be confirmed during the analysis. Lastly spraying more than 10 days before the first appearance of late blight is assumed to have no significant effect, hence is considered to be the optimal day of fungicidal application.

Chapter 2

Method

Attacking the questions at issue there are several possible methods which may be considered. The large and vaguely defined issue regarding data analysis its predictability was investigated using multiple linear regression. Predicting the appearance day of late blight was made using Cox regression, SIMCAST and a continuous variant of SIMCAST [8]. The Cox models was compared using Akaike's information criterion to find the most suitable covariates and then validated using leave-one-out cross-validation.

2.1 Multiple linear regression

Given a response variable **Y** and p covariates **X**, one may define a model as:

$$Y = X\beta + \epsilon \tag{2.1}$$

given n observation it may be written in the following form:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,p} \\ x_{2,1} & x_{2,2} & \dots & x_{2,p} \\ & & \vdots & \\ x_{n,1} & x_{n,2} & \dots & x_{n,p} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_p \end{bmatrix} \tag{2.2}$$

Assuming independent identically distributed residuals, $\epsilon_i \sim N(0, \sigma^2)$, and viability of 2.1 the best estimates in least square sense is given by:

$$E[\hat{Y}|X] = X\hat{\beta} = X(X^TX)^{-1}X^TY \tag{2.3}$$

Then the parameter variance is given by:

$$Var(\hat{\beta}) = (X^TX)^{-1}\sigma^2\epsilon \tag{2.4}$$

Notable is one must be cautions regarding correlated covariates etc. Also more results such variance etc. may be derived.

2.2 Cox regression

Survival analysis has some characteristics due to its nature. It aims to predict the risk of an event given it has not yet occurred at some time t, this is know as right censoring [9]. Left censoring is also possible to include in the model, i.e. adding a criterion which accounts for contaminated samples at the initial time [9]. Neither linear nor logistic regression are able to model the mentioned property hence survival analysis is a valuable tool [9].

The survival function at time t is defined as:

$$S(t) = 1 - F(t) \tag{2.5}$$

where $F(t)$ denotes cumulative event function, i.e. the probability of an event occurring given it has not happened yet [9]. However the short-term event rate of risk for a sample prior to exposure, also know as the hazard function, $h(t|x)$, may also be of interest [9]. A more proper definition of the hazard function is:

"The hazard function h(t) is the short-term event rate for subjects who have not yet experienced the outcome event."[9]

Similarly the corresponding hazard ratio is defined by

$$HR(t) = \frac{h(t|x)}{h_0(t)} \tag{2.6}$$

where $h_0(t)$ is some reference group[9]. The Cox-regression assumes log-linearity of the hazard function hence rewriting with respect to the hazard function the model is obtained:

$$h(t|x) = h_0(t)e^{\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p} = h_0(t)HR(t) \tag{2.7}$$

where $X = [x_1, x_2, \ldots, x_p]$ are the covariates [9] and $\beta = [\beta_1, \beta_2, \ldots, \beta_p]$ [9] are the parameters[1]. The reference $h_0(t)$ may also be interpreted as an intercept. Note that $HR(t) > 0$ follows from the log-linearity, moreover the hazard function is assumed to be constant [9]. Comparison of Cox models may be done using likelihood ratio tests [9]. Note that in parametric models it is assumed that $h_0(t)$ has some given distribution where as a semi-parametric Cox regression does not[9]. The latter mentioned is considered to be more robust due to the adaptability[9].

[1]Note the similarity to linear regression

2.3 SIMCAST

Over the time several models have come up trying to predict the appearance of late blight, of which one is SIMCAST. Due to its simplicity and relative efficiency it has been popular. Moreover several modifications to the original method has been made improving the results.

Its input is number of consecutive hours where relative humidity above 90% during one day, and the daily mean temperature [10]. Depending on these so called blight units are calculated and cumulatively summed up, when those reach a set threshold it is recommended to start treating against the late blight.
A continuous representation of SIMCAST was suggested by *Veiko Lehsten* [8], who fitted a logistic function (s) with respect to time (t) interval where relative humidity was above 90% and a beta function (β) with respect to mean temperature T:

$$s(t) = \frac{a}{1 + e^{-(b+ct)}} \tag{2.8}$$

where variables calibrated to the given case are $a = 7.325$, $b = -0.5270$ respectively $c = 4.6006$.

$$\alpha = \frac{\log(2)}{\log\left(\frac{T_M - T_m}{T_o - T_m}\right)} \tag{2.9}$$

$$\beta(T) = \frac{2(T - T_m)^\alpha (T_o - T_m)^\alpha - (T - T_m)^{2\alpha}}{(T_o - T_m)^{2\alpha}} \tag{2.10}$$

where $T_m = -1.9453$ is the minimal temperature for growth, $T_o = 17.1268$ the optimal and $T_M = 29.6298$ the maximal. This continuous version of SIMCAST will also be referred to as cSIMCAST

Algorithms, in R code, for both variants of SIMCAST will be supplied in appendix 5.

The SIMCAST is based on the following lookup table:

		Consecutive hours of RH ≥ 90% that should result in blight units of:							
Avg. temp (°C)[b]	**Cultivar resistance[c]**	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
>27	S	24	…	…	…	…	…	…	…
	MS	24	…	…	…	…	…	…	…
	MR/R/HR	24	…	…	…	…	…	…	…
23–27	S	6	7–9	10–12	13–15	16–18	19–24	…	…
	MS	9	10–18	19–24	…	…	…	…	…
	MR/R/HR	15	16–24	…	…	…	…	…	…
13–22	S	6	…	…	…	…	7–9	10–12	13–24
	MS	6	7	8	9	10	11–12	13–24	…
	MR/R/HR	6	7	8	9	10–12	13–24	…	…
8–12	S	6	7	8–9	10	11–12	13–15	16–24	…
	MS	6	7–9	10–12	13–15	16–18	19–24	…	…
	MR/R/HR	9	10–12	13–15	16–24	…	…	…	…
3–7	S	9	10–12	13–15	16–18	19–24	…	…	…
	MS	12	13–24	…	…	…	…	…	…
	MR/R/HR	18	19–24	…	…	…	…	…	…
<3	S	24	…	…	…	…	…	…	…
	MS	24	…	…	…	…	…	…	…
	MR/R/HR	24	…	…	…	…	…	…	…

[a] Blight units are estimated during a 24-h period using hourly observations starting at 1300 hours and ending at 1200 hours the following day. These rules were modified from Fry et al. (5) to work for Mexican cultivars under tropical highland conditions
[b] Average temperature during hours of RH ≥ 90%.
[c] Level of resistance of a potato cultivar to late blight: S = susceptible, MS = moderately susceptible, MR = moderately resistant, R = resistant, HR = highly resistant.

Figure 2.1: SIMCAST look up table [10].

2.4 Akaike's information criterion

In order to compare statistical objects with different covariates based on the same data where the likelihood is available there are several methods. One widely used is Akaike's information criterion defined as:

$$AIC = 2k - 2\ln(L(\hat{\beta}|y)) \tag{2.11}$$

where $L(\hat{\beta}|y)$ is the maximized likelihood function of the model and k the number degrees of freedom. Choosing between models the one related to the lowest AIC is preferable.

2.5 Imputation

It is not rare that data sets in the real world have missing data. The reasons are many, from time efficiency which may be intentional to changing experiments programmes etc. As it is a common issue, also many methods have been developed to cope with such problems, of which one is imputation.

Imputation estimates the complete quantity (Q), by filling in unbiased missing samples, i.e.

$$E[\hat{Q}|Y] = Q \tag{2.12}$$

such that its confidence valid. This means the estimates must be such they do not decrease the variance and thus still can be used in test statistics [11]. There are several different imputation methods of which predictive mean matching was used. Predictive mean matching starts by linearly estimates missing samples [12]. Then a subset of observations close to the estimate for respective sample is defined [12]. Then a random draw from these will be made which will be used as the imputed value [12].

2.6 Elastic Net

Elastic net is one way to estimate covariate parameters $\hat{\beta}$, i.e.:

$$\hat{\beta} = \text{argmin}_{\beta} \left(||Y - X\beta||_2^2 + \lambda(\alpha||\beta||_1 + (1-\alpha)||\beta||_2^2) \right)$$

where λ is amount penalization adding a covariate, and α the weight between the one- and two-norm penalty [13]. If $\alpha = 0$ this is know as ridge regression and with $\alpha = 1$ it called lasso [13]. If there are heavily correlated covariates elastic net tends to choose only a few of these, it also preforms more comparable to ridge regression (which tends to most covariates) [13].

2.7 Cross-validation

There are multiple ways to validate models, of which is preferable may depend on the data sample size etc. It is not uncommon to divide the data to three sets, training, testing and validation [14]. However by such methodology data is disregarded in the modeling stage as parts is only used in the validation stage which may be vital concerning small data sets, moreover choosing the sets may not be trivial [14]. To overcome some of the issues cross-validation may be an option to evaluate the prediction errors and robustness of models [14].

Cross-validation is a family of methods comprising k-fold and leave one out validation etc. of which also there are sub-groups [15]. Leave one out removes a single observation from the modeling data which is then used as validation [15]. Iterating over all observations, i.e. first do the processes removing observation (1), then observation (2) and so on [15]. Then suitable measures may be defined from the combined diagnostics such as mean standard deviation, success ratio etc [14].

As the model of interests may be updated yearly also leaving one year out of the set should be preferable in a validation stage hence leave one out will be used. Measures which will be considered is:

Mean standard deviation

Mean failure rate

Mean accurate prediction[2]

Note that the complement to failure rate is success ratio hence was omitted in the analysis.

[2]This will be defined further on.

Chapter 3

Results

There are several parts within the analysis. First an overview of the data and characterizations will be considered. Then investigation using a survival approach in order to deviate the effect of treatments followed by modeling of yield in order to consider the insurance.

3.1 Data

Since 1983 an experiment comparing different fungicidal treatment programs has been issued in three stations cultivating Bintje potatoes. In order to compare these fairly, the effects of other variables have been reduced according to recommendations, example wise pest control, fertilization and irrigation. Due to varying regulations, technological advancement and research results these together with the treatments also varies over time. As weather related data was missing during the trials reconstructed values based on historical observations was used.

3.1.1 Potato cultivation data

Yearly data relative planting date collected at the following locations: Kristianstad Mosslunda C_1, Bjärred Borgeby C_2 and Lilla Böslid Eldsberga C_3 [2].

Figure 3.1: Kristianstad Mosslunda C_1, Borgeby C_2 and Lilla Eldsberga C_3.The following image was plotted using Google maps.

Each station measured several variables, those are presented in table (3.1)

Table 3.1: Cultivation data variables collected from experimental data.

Variable	Unit	Notation
Planting day	Days since 1^{st} Jan	Pd
Harvest day	Days since 1^{st} Jan	Hd
First fungicidal application	Days since 1^{st} Jan	Fa
Last fungicidal application	Days since 1^{st} Jan	La
Number of treatments		nT
Non-treated		T_1
Best treatment	based on yearly results	T_2
Control group of best treatment	based on yearly results	T_3
Worst treatment	based on yearly results	T_4
Chemical defoliation day	Days since 1^{st} Jan	Dd
Organic matter content	%	OrgC
pH	%	pH
Clay content	%	ClayC
Nitrogen applied		N
Phosphorus applied		P
Potassium applied		K
Amount of irrigation	mm	Irr
Number of irrigations		nIrr
Estimated late blight severity	%	LBs
Yield	ton/ha	Yield
Measures of fertilization need for different substances		PAL, KAL, CaAl, MgAl and KMgQ

Need of fertilization was preserved using KAL and PAL scales [2][16]. Some data, especially in the early stages is missing. These was reconstructed using the imputation module in R[1]. The affected variables was: Appearance of late blight for treated experiments (early years only), organic matter content (OrgC), clay content (ClayC), pH, fertilization measures (PAL, KAL, CaAl, MgAl, KMgQ), amount and number of irrigation (Irr, nIrr).

[1]Predictive mean matching (pmm) methodology was used whilst imputing.

3.1.2 Weather data

As local weather data was not originally obtained an estimated data set was used. This was done by *Tinghai and Deliang in Gothemburg* who used TAPM version 4 for corresponding station $(C_1, C_2, C_3,)$. The following data was included:

Table 3.2: TAPM estimated weather data containing time, wind seed, temperature, rain fall, relative humidity. Using the relative humidity also Relative Humidity ≥ 90 was evaluated.

Variable	Unit	Notation
Time	h	t
Wind speed	m/s	Ws
Temperature	^{o}C	T
Rain fall	mm/h	Ra
Relative Humidity	%	RH
Relative Humidity ≥ 90	%	rH

3.1.3 Data transformations

Assuming the *Phytophthora infestans* do only affect and develop when the plants has emerged from the ground it is preferable to use this date. However in the original data set such information is missing. One way to evaluate emergence of potatoes is using GDD growing degree day defined as [17]:

$$GDD = \frac{T_{max} + T_{min}}{2} - T_b \tag{3.1}$$

where T_{max} and T_min are the maximal- respectively minimal daily temperature. T_b is a constant baseline related to species and its growth properties. This baseline is estimated to be 2^oC for Nordic countries [8].Then the emergence day (Ed) is estimated by summing GDD from planting day until the summation level 220 is reached or exceeded, i.e.:

$$Ed = Pd + min_n \left(\sum_{n=Pd+1}^{\infty} \mathbb{I} \left((\sum_{i=Pd}^{n} GDD_i) \leq 220 \right) \right) \tag{3.2}$$

Notably is that the model only considers temperature disregarding factors such as sun etc.

Due to difference in time scale between the yearly samples was replicated such it was of the same length as the hourly. Also a set downscaling to daily means of the weather data was considered in similar manner. Beyond above functions also cumulative sums, starting at the emergence date, of the data was considered (in corresponding time scale). Examples of such, for i from emergence to current day n, is:

Table 3.3: Example cumulative sums considered as covariates. The following notations are used: Temperature (T), time (t) and relative humidity. The functions $s()$ and $\beta()$ are defined as in equation (2.8) respectively (2.10).

$T_n^{(c)}$	$\sum_{i=1}^{n} T_i$
$\beta(T)_n^{(c)}$	$\sum_{i=1}^{n} \beta(T_i)$
$t_n^{(c)}$	$\sum_{i=1}^{n} t_i$
$s(t)_n^{(c)}$	$\sum_{i=1}^{n} s(t_i)$
$rH_n^{(c)}$	$\sum_{i=1}^{n} rH_i$
$(\beta(T)\ rH)_n^{(c)}$	$\sum_{i=1}^{n} rH_i\ \beta(T_i)$
$(\beta(T)\ s(t)_n^{(c)}$	$\sum_{i=1}^{n} s(t_i)\ \beta(T_i)$
$R_n^{(c)}$	$\sum_{i=1}^{n} R_i$

As there are several time scales in the different data sets it adds possibilities on how and when to apply different functions. Note this matters as the considered functions are not commutative. Generally daily maximums and means was considered and then relevant functions applied as it suggests above. One more covariate worth mentioning, which is similar to one above, is applying the constitutions SIMCAST for hourly samples starting at plating date and then taking the daily mean (instead of daily mean followed by cSIMCAST). I.e. let

$$\begin{cases} b_0 = 0 \\ b_h = b_{h-1} + \beta(T_h)s(t_h)/24 \end{cases}$$

where the temperature T_h and time interval where relative humidity is above 90% is hourly sampled from the planting day ($h = 0$), then $\delta SIMCAST$ is the daily mean from the cumulative sum b_h

$$\delta SIMCAST_n = \sum_{h=Pd+n*24}^{Pd+(n+1)*24} b_h/24. \tag{3.3}$$

3.2 Regression and illustrative plots

As previous section suggested there are lots of different covariates available which individually are assumed to affect or neutralize effects the experiments in some manner. There are however some main ingredients to the analysis i.e. yield, appearance of late blight, treatments, regions and a time aspect. Briefly looking at the data a suspicion arise concerning both the yield and late blight as 1994 the characteristics seem to change:

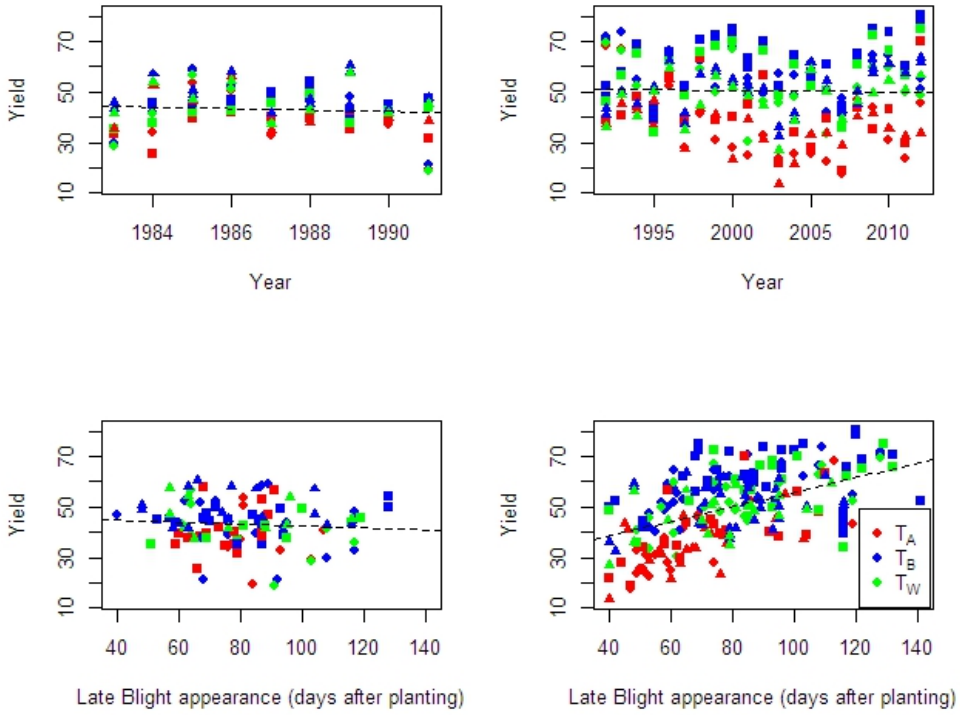


Figure 3.2: Top image is yield plotted as a time series with its trend. Bellow the yield with respect to late blight appearance (counted from planting day) is illustrated. Where T_A is sample where no treatment was given and T_B the best sample with respect yield and T_W the worst with treatment. Shapes corresponds to the three different counties.

The yield variance and mean yield looks increased after the change-point 1994, see Figure (3.2). Moreover the effect of late blight, with respect to infection day after planting, had no clear correlation in the first observed years where it is as strong after 1994. Typically the treatments are effective and increase the yield and also delay the infection rate, see the clustering in Figure(3.2).

Using linear regression with respect to appearance date of the late blight:

Table 3.4: Regression models with respect to year respectively appearance date of late blight. All presented parameters are significant at 5% significance level.

Model	β_0	β_1	Evaluated for the years
$y_n = \beta_0 + LB_n\beta_1$	36.1 ± 3.8	0.11 ± 0.04	1983-1993
$y_n = \beta_0 + LB_n\beta_1$	23.9 ± 2.9	0.33 ± 0.04	1994-2012

Table (3.4) confirm that there is a significant difference of late blight appearance around 1994, both in effect and mean. This may interpreted as the dynamics change around 1994. During both periods considered, a significant yearly effect was found, moreover using year and appearance as covariates during 1993-1994 only year was significant[2]. Hence further analysis considers only the period 1994-2012.

[2]This assumes a 5% significance level

As there is large variance in the data from unknown sources, differences between treated and non-treated experiments was considered. In practice it would be possible to use a reference assuming its similar in cultivation conditions and thereby increasing the robustness. Note such difference would correspond to prevented loss in yield caused by late blight. There are differences between the regions, and also it seems to be some correlation as well. Measuring using Pearson's correlation:

Table 3.5: Pearson's correlation between stations with respect to the difference between treated and non-treated experiment respectively only treatment, i.e. comparing yield between stations using the same treatment.

Counties	Treatment - non-treated	Correlation	Treatment	Correlation
C_1 vs C_2			T_1	0.44
C_1 vs C_2	$T_2 - T_1$	0.61	T_2	0.37
C_1 vs C_2	$T_3 - T_1$	0.72	T_3	0.31
C_1 vs C_2	$T_4 - T_1$	0.64	T_4	0.40
C_1 vs C_3			T_1	0.60
C_1 vs C_3	$T_2 - T_1$	0.63	T_2	0.38
C_1 vs C_3	$T_3 - T_1$	0.70	T_3	0.22
C_1 vs C_3	$T_4 - T_1$	0.68	T_4	0.23
C_2 vs C_3			T_1	0.45
C_2 vs C_3	$T_2 - T_1$	0.54	T_2	0.52
C_2 vs C_3	$T_3 - T_1$	0.51	T_3	0.58
C_2 vs C_3	$T_4 - T_1$	0.46	T_4	0.40

Table (3.5) also suggest there are strong correlation between the counties in particular regarding the relative difference.

Translating yield into some currency allows comparison of profit with respect to the number of fungicidal applications. According to the authors of *Overview and Breeding Strategies of Table Potato Production in Sweden and the Fennoscandian Region* [18], a reasonable estimate of the one spray is 425 SEK/ha. Whereas the expected yield is approximately worth 990 SEK/ton [19]. These estimates will be used through out the analysis.

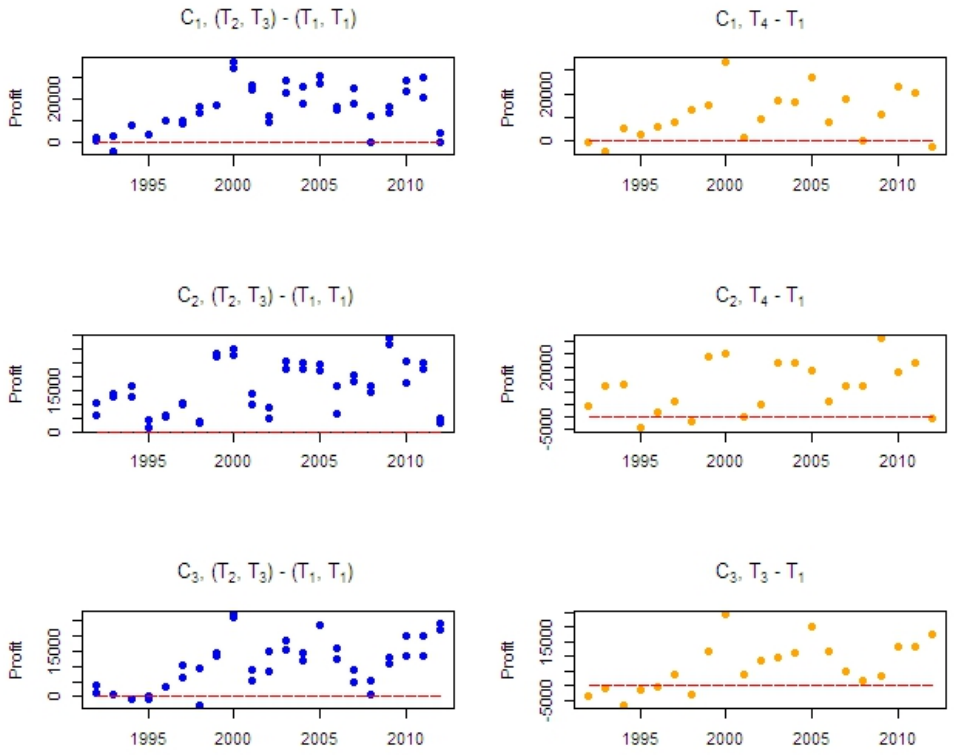


Figure 3.3: Estimated relative profit of spraying per ha for respective county and treatment over time. Example wise, $C-1, T_4-T_1$ denotes in county C_1 the difference between untreated experiment and treated with T_4, is used to calculate the profit gained by using fungicides. In the left plots vector notation is used as T_2 and T_3 uses the same treatment and thereby are plotted together.

Generally spraying is preferable increasing the profit, see dots above the line in Figure (3.3). This amount depend on the treatment and more over there is variation within the same treatment.

3.3 SIMCAST

A famous method used to predict late blight is as previously mentioned, SIMCAST. *Lehsten et al* continuous representation was strictly increasing due to the nature of logistic functions which has a large effect when considering the cumulative sums [8]. The continuous SIMCAST is more stable concerning growth in blight units, hence the results are expected to deviate slightly. The following cumulative blight units would be given concerning for considered experiments:

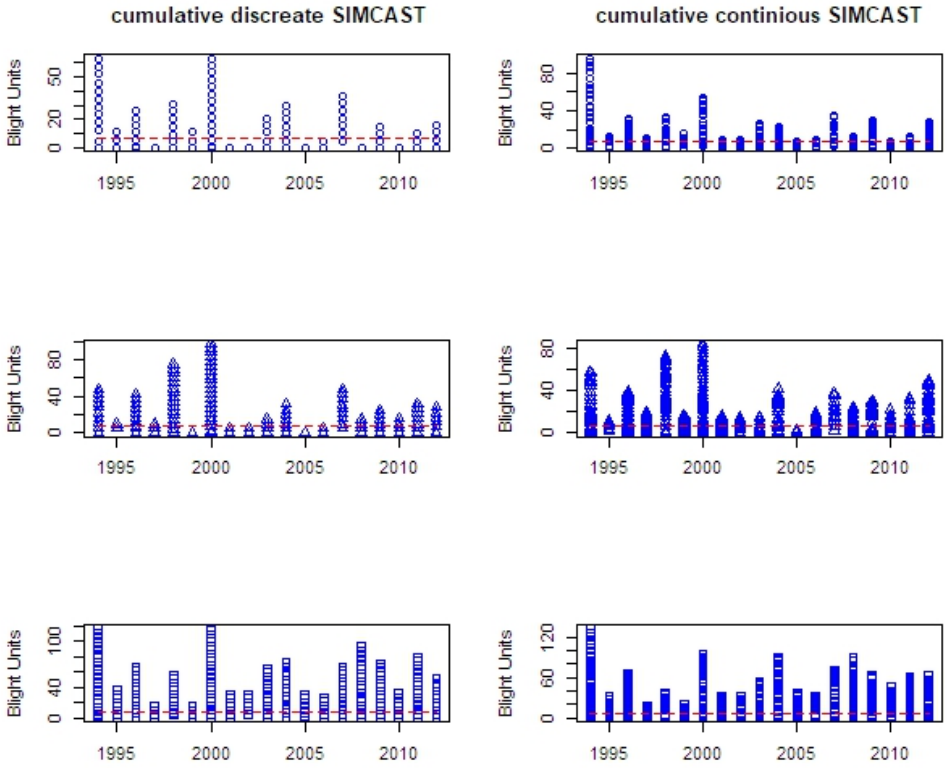


Figure 3.4: Discrete SIMCAST models to the left, cSIMCAST to the right corresponding to C_1, C_2 and C_3 from the top and down. A threshold (red dashed line) at 7 are set in the discrete case also 7 in the continuous, at which it would be recommended to start application of fungicides. All observations are present each year i.e. the highest obtained yearly blight units level is when the appearance occurred.

Thresholds was chosen such SIMCAST would not fail with approximately 5% accuracy, i.e. set to 7 blight units. Both SIMCAST and its continuous

representation gives similar results, however not identical, see Figure (3.4). Note that the discrete case tends to move less which is due to generating more zeroes (for time intervals up to 6 hours), where as the continuous only do so when the time interval is less than one. Most often SIMCAST recommend spraying prematurely, which can be seen by the amount of samples above the threshold in (3.4). In spite of this it does fail not to rarely, see years where the cumulative blight units does not reach the threshold.

In order to get a better visualization three years in C_1 is presented bellow:

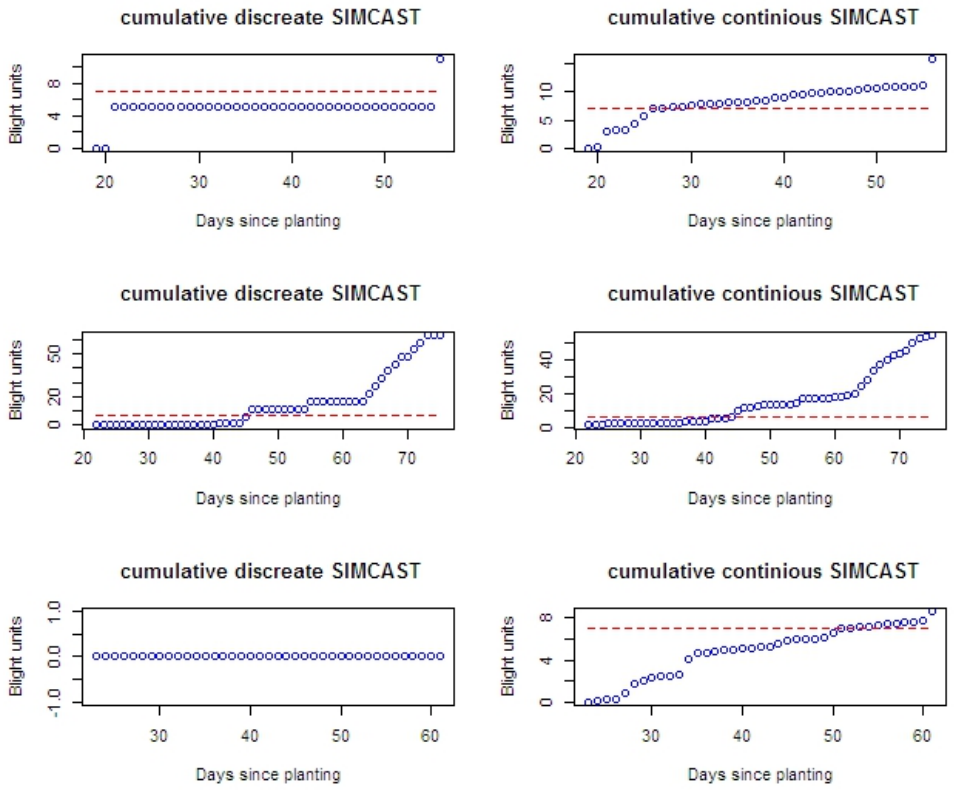


Figure 3.5: Discrete SIMCAST models to the left, cSIMCAST to the right corresponding to the years 1999, 2000 and 2001 for site C_1, Kristianstad, (receptive year in top left image in (3.4)). The same thresholds as in Figure (3.4) are used, i.e. seven blight units (marked as red dashed line).

The classic SIMCAST has issues due to lack of a minimum growth and does either not reach the threshold or overcompensate heavily, see Figure (3.5). Note that bottom left image in (3.5) does not even have any blight units. The continuous representation has a growth property which seems to be beneficial finding the occurrence more accurately, in particular during dry years. However it also heavily overcompensate if it the main emphasis is to succeed in prediction. Calculating failure rate for cSIMCAST with respect to optimal threshold level it will increase linearly, similarly predictions within 10 days interval to the event is convex with maximum around 25 blight units.

3.4 Cox Regression

As previously mentioned applying survival analysis for modeling additional possibilities arise such as accounting for both left and right censoring. As the emergence day is approximated its reasonable to believe that this should be accounted for (left censoring) as the disease might have started before the estimated date. As for right censoring it is the attribute which is of interest as the exact time of late blight occurrence is not known. Note that estimating the event only non-treated data was used for the general model due to the delaying effect of treatments.

By considering survival analysis two things may be investigated, whether the treatment has a significant effect delaying the late blight and also finding a model which may be used for predictions of attacks. Due to the large variety of covariates also several possible models arise, however correlation between those reduce the number of valid possibilities. Considering both variables which remain relatively constant each year, i.e. cultivation related data, and weather data with low correlation between covariates several models was tested. All single covariate models was compared using their log-likelihood. Several other transformations and similar was investigated such as applying functions on hourly samples before using mean etc.

Note that the appearance of late blight, i.e. the event, and time is common in all above cases hence only the covariates are presented. The best of these was $\delta SIMCAST$ i.e. daily mean of cumulative cSIMCAST based on hourly samples togeather with planting day and PAL as covariates. Second best was using the $(\beta(T)rH)_n^{(c)}$, which is the cumulative sum of the daily mean of $\beta(T)$ times the relative humidity above 90%. Using the SIMCAST blight units as covariates both are significant with 5% significance level but worse than previously mentioned. Worth mentioning the continuous SIMCAST served worse as a covariate than the discrete counter part.

Starting considering $(\beta(T)rH)_n^{(c)}$, manual forward-backward selection was made cautiously regarding correlated covariates. In each step the AIC was evaluated when adding or removing a covariate, also the first order intersection was included in the covariates. Including year as a yearly constant and station (C_1, C_2, C_3) as a categorical gave:

Table 3.6: Coefficients of Cox $model_1$. I.e. parameters used in Cox regression to predict late blight appearance date.

Covariate	Coefficient	Standard error	p-value
$(\beta(T)rH)_n^{(c)}$	0.0026500	0.0006416	0.0000362
C_2	-0.4838165	0.3709108	0.1921
C_3	-0.9000176	0.3720256	0.0156
Year	0.0684976	0.0307927	0.0261

However changing reference county to C_2 one obtain:

Table 3.7: Coefficients of Cox $model_2$.

Covariate	Coefficient	Standard error	p-value
$(\beta(T)rH)_n^{(c)}$	0.002650046	0.0006415797	0.0000362
C_3	-0.416201108	0.3675532688	0.2574848
C_1	0.483816505	0.3709107755	0.1920968
Year	0.068497632	0.0307927111	0.02611647

Table (3.6) suggests there is a significant difference between C_1 and C_3, however Table (3.7) states there is no significant difference between C_2 and C_1 nor C_3 and C_1. All stations will remain in the model due to it may be of further interest. Note that generally the infection starts in Lilla Eldsberga (C_3) first, Borgeby (C_2) secondary and in Kristianstad (C_1) last, see (3.7).

Adding all treatments to evaluate their effect in appearance date similarly to previous suggests there is no significant difference between T_1 and T_2 where as T_2, T_3 and T_4 are significantly different. Using stratification to illustrate (disregarding counties):

Both the categorical analysis and Figure (3.6) suggest that treatments may have a positive effect delaying the infection date, however bad treatments does not affect the appearance date significantly (by regression it can be shown it still affects the yield).

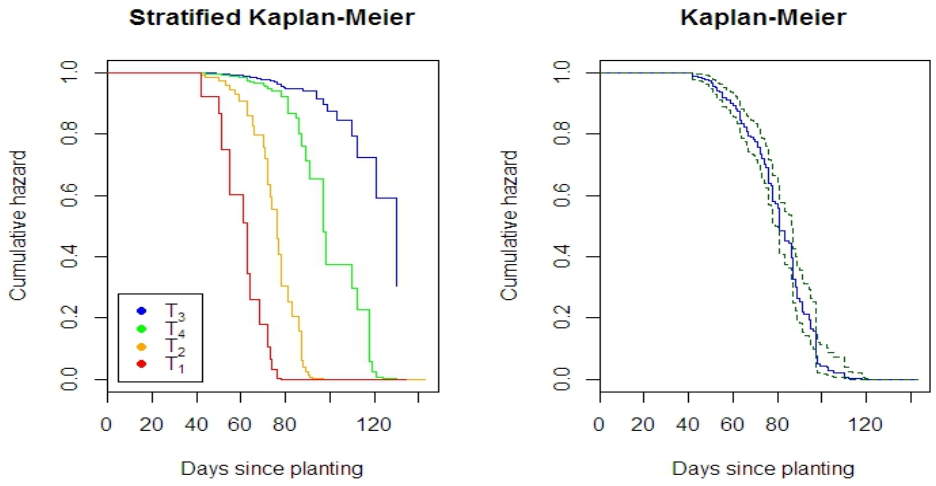


Figure 3.6: Kaplan-Meier, cumulative hazard with respect to days after planting. Stratified with respect to treatment to the left, and categorical with 5% confidence interval to the right.

Similarly considering $\delta SIMCAST$ the best model found was using it together with PAL and Pd. Notable is that the influence from county is totally insignificant in this case. Also a possible choice of covariates is replacing PAL and Pd with year obtaining:

Table 3.8: Coefficients of Cox $model_3$.

Covariate	Coefficient	Standard error	p-value
$\delta SIMCAST$	0.146453	0.020715	10^{-11}
PAL	0.062754	0.009261	10^{-10}
Pd	0.048022	0.010607	10^{-5}

respectively:

Table 3.9: Coefficients of Cox $model_4$.

Covariate	Coefficient	Standard error	p-value
$\delta SIMCAST$	0.13901	0.01983	10^{-11}
Year	0.04030	0.01159	0.0005

Comparing all models using AIC the best Cox model is given in table (3.8) and the second best in (3.9).

3.5 Predictions

Assuming all weather data is valid to use as predicted data SIMCAST, contentious SIMCAST, foliage connection and the proportional Cox regression models may be used to predict the appearance date of late blight.

3.5.1 SIMCAST, cSIMCAST and foliage prediction performance

Using SIMCAST a threshold which when reached suggest application of fungicides. Striving after a failure rate which is less than 5% the threshold must be set low, see Figure (3.4). Note the years where late blight appearance is correlated to the yield is used, i.e. 1994-2012. Using a threshold of 7 blight units gives an empirical failure rate of 20%. The such rate possible obtained, i.e. assuming a threshold of 1, is 13%. Moreover is the variability large from failure to 90 days ahead prediction.

Using the cSIMCAST further flexibility is added due to not as restricted growth criterion hence it is possible to fit an empirical threshold more accurately obtaining 5% failure rate. The accuracy in prediction date is however slightly worse and also has a larger variability than corresponding in SIMCAST.

During experiments the first application of fungicides was connected to the foliage progress, when it met between rows was the application criterion, this will be referred to as foliage prediction. Such method gave 18% failure rate but had better accuracy with the largest miss prediction at 66 days and also a lower variance.

On the following page predictions are plotted marking good attempts, successful but overcompensated and failures.

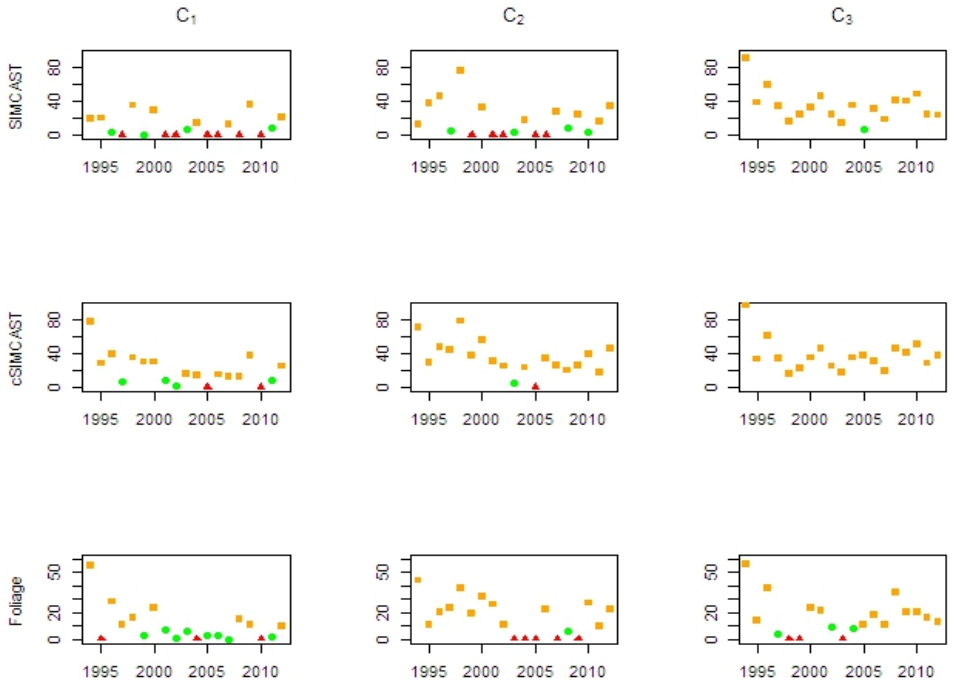
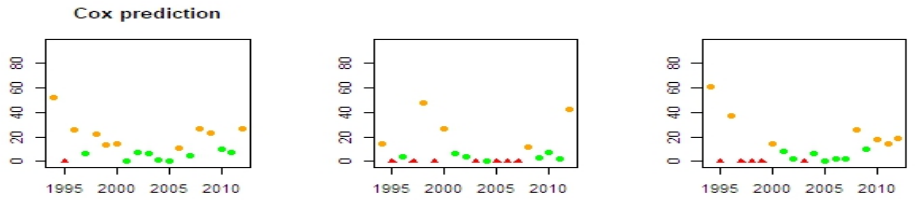


Figure 3.7: Forecast of late blight in the counties C_1, C_2 and C_3 in given order from the left to the right. In top row SIMCAST was used, second continuous and in bottom spraying as in experiments i.e. when foliage met between rows. Triangles denote failure, i.e. not able to predict event 10 days ahead. Circles denotes successful attempts within the interval 10-20 days prediction (hence considered to be good predictions). Squares are attempts which was sprayed ahead, but overcompensated. The Y-axis denoted number of days until late blight appearance. For consistency both SIMCAST implementations used a threshold of 7 blight units in these plots.

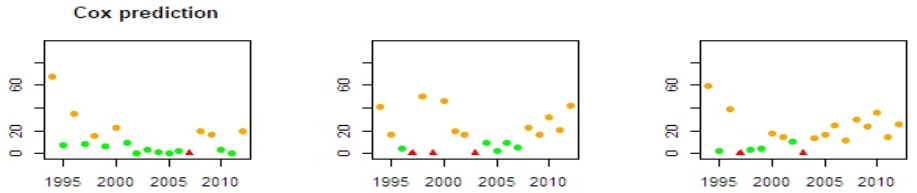
Comparing these methods only cSIMCAST can be used with 5% failure rate, note this is empirical results with thresholds adjusted to fit the data, moreover such low threshold would increase amounts of sprays. Around year 2000 all predictions seem to stabilize with a lower variance, except for SIMCAST in C_1 which already was at a stable state, see Figure (3.7). Foliage prediction and SIMCAST preforms approximately the same in failure rate, where as SIMCAST is worse regarding the variance. Also worth noting is neither method performs accurately, i.e. few which succeed to predict within 10 days. There also seem to be different success between counties. Including data during 1983-1992 does not add any additional failures moreover they are similar to the first three years concerning variance etc. The continuous SIMCAST also tend overcompensating due to its enhanced growth ability.

3.5.2 Cox prediction

In order to compare Cox regression to previous methods, predictions was made for each year. The data was adjusted by iteratively removing samples of the year which was to be predicted redefining the model for each prediction, hence lowering the influence of year which was of interest, i.e prediction given by leave one out cross validation. Using covariates given in (3.7) and similarly but stratifying with respect to station:



(a)



(b)

Figure 3.8: Prognosis of late blight in the counties C_1, C_2 and C_3 in given order from the left to the right using Cox prediction. Red triangles denote failure, i.e. not able to predict event 10 days ahead. Green dots circled by red denotes attempts which was sprayed ahead but was overcompensated. The Y-axis denoted number of days until late blight appearance. Threshold of expected hazard ratio is set to 5%. (a) is prediction using station as a categorical variable and (b) stratifying with respect to station.

Using the expected hazard ratio lower than 5% gives 10% failure rate with a standard deviation in predicted day at 16 days considering the same base line hazard (using categorical variables). Moreover 33% of the prediction was within a preferable time, i.e. not further than 10 days ahead. Assuming the counties does not have a common base hazard line, i.e. a stratified model with respect to station would give 23% failure rate with 15 days in standard deviation and 37% accurate enough predictions.

Predicting with thresholds between 0.001 and 0.5, using (3.6), the failure rate increased linearly whereas the accurate prediction rate had is has a maximum at the threshold 0.05. Note that it is possible to have more restrictive strategies using a lower threshold hence indication of first application would come earlier.

3.6 Elastic net cox regression

To find a suitable penalty weight (α),

$$\alpha = [0, 0.1, 0.25, 0.4, 0.5, 0.6, 0.75, 0.9, 1]$$

was considered. Evaluating these Cox models using the R module *glmnet* [20][21] no convergence of the penalty λ was found hence the minimum $\lambda = 0.0025$ was used.

Table 3.10: Overview of elastic net Cox regression models using different weights α and $\lambda = 0.0025$. Failure (F) is non-successful attempt and an accurate prediction (AP) is a successful attempt where the event occurred within 10 days of the prediction. The standard deviation Std is also given in days.

α	0	0.1	0.25	0.4	0.5	0.6	0.75	0.9	1
F	28%	30%	30%	28%	30%	30%	30%	26%	32%
AP	32%	32%	30%	30%	28%	28%	28%	30%	26%
Std	12	12	13	13	14	13	13	14	13

To evaluate the behavior with respect to the penalty λ, the failure ratio, amount of accurate predictions and standard deviation (in days) was calculated for the grid:

$$\lambda = 10^{-2}\ [0.1, 0.8, 1.5, 2.2, 2.9, 3.6, 4.3, 5.0, 5.7, 6.4, 7.1, 7.8, 8.5, 9.2, 9.9]$$

and the same α as before. The results[3] suggested that with respect to standard deviation it is preferable to choose a small λ (this is the measure *glmnet* [20][21] uses). The choice of λ does not affect the other measures very much. Accurate predictions with respect to α suggests that $\alpha = 0$ (ridge regression) is the best. Failure ratio and standard deviation is also better for smaller α.

The best elastic net model was obtained using $\alpha = 0$ and $\lambda = 0.022$. The related Cox regression model will be referred to as *Ridge Cox*[4]. The best non-ridge regression model was obtained with $\alpha = 0.6$ and $\lambda = 0.015$. Its related Cox regression will be referred to as *Elastic Cox*. More details about accuracy, failures and standard deviation is given in the next section Table (3.11). Notably is that the best parameters according to an elastic net Cox regression model using $\alpha = 0.5$ (equal mix of one- and two-norm penalties) is given using $\delta SIMCAST$, planting day, estimated emergence day and time since planting.

[3] All results will not be presented but only the best.

[4] This due to using given parametrization it is a ridge regression.

3.7 Prognosis overview

Using the different models neither SIMCAST nor cSIMCAST was successfully in accurate predictions and obtaining low failure rate, see Table (3.11). Also notable is both those also have an adjusted threshold to ensure a reasonable failure rate. Stratified Cox assumes different models for each station and is not likely to be usable in practice but still preformed rather well.

The best Cox model based on failure rate (non-stratified) was Cox $\delta SIMCAST$ with parameters given in Table (3.8), followed by those presented in Table (3.6), respectively (3.9) and lastly a Cox model using only $(\beta(T)rH)_n^{(c)}$ and year as covariates. Best with respect to accurate predictions and standard deviation are similar, see Table (3.11).

- Failure,
  model failed to predict late appearance ten days in advance
- Accurate prediction,
  successful prediction where event occurred within ten days
- Standard deviation (in days)

Summarizing the prognostics:

Table 3.11: Overview considered prognosis models. All models assume the same thresholds except $cSIMCAST^{(*)}$ which has a threshold of 15 blight units. Note that failures was excluded evaluating accurate predictions and standard deviation due to if failure occurs it is assumed fungicidal application starts immediately.

Model	Failures	Accurate predictions	Standard deviation (days)
SIMCAST	20%	15%	18
cSIMCAST	5%	8%	19
$cSIMCAST^{(*)}$	22%	20%	19
Foliage	21%	23%	13
Stratified Cox	11%	33%	16
Categorical Cox	23%	37%	15
Cox (disregarding C_i)	32%	25%	14
Cox $\delta SIMCAST$(3.8)	22%	42%	9
Cox $\delta SIMCAST$(3.9)	25%	35%	12
Ridge Cox	25%	26%	15
Elastic Cox	23%	30%	14

3.8 Profit using alternative models

Assuming the cost is relative the prices in 2014, the yield is approximately worth 990 sek/ton and each fungicidal application costs about 425 sek/ha[18][19]. Moreover spraying 10 days ahead gives optimal effect and by lowering the number of applications also reduce the total cost implying increased profit, recall section (1.2.1). Using conducted experiments given the difference in excess spraying cost using cSIMCAST, foliage prediction respectively Cox model(3.7):

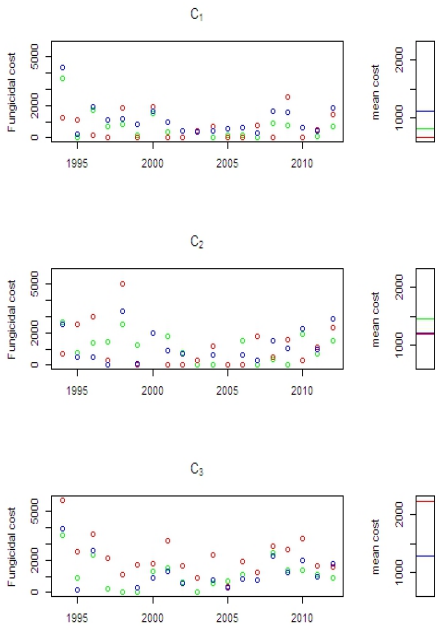


Figure 3.9: To the left: Yearly excess spraying cost assuming constant application cost 425 sek and application should be made at least 10 days ahead of late blight appearance. Red corresponds to using cSIMCAST, green Foliage prediction and blue Cox model (3.7), i.e. considering station as a categorical variable. To the right: Mean excess cost of successful attempts using the same colour codes as previous, note failures are disregarded in this mean.

Disregarding all failures cSIMCAST is most costly followed by foliage prediction and the best was Cox model(3.7), see Figure(3.9).

Similarly comparing categorical and stratified Cox models with respect to station respectively disregarding the station:

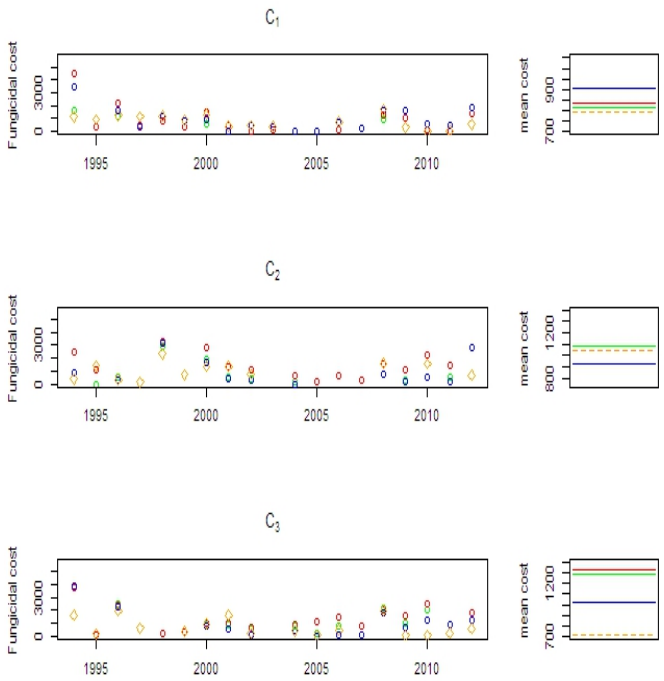


Figure 3.10: To the left: Yearly excess spraying cost assuming constant application cost 425 sek and coating should be made at least 10 days ahead of late blight appearance. Red corresponds to prediction using Cox model with station as a categorical variable, red a stratified model with respect to station and green disregarding station variables. To the right: Mean excess cost of successful attempts using the same colour codes as previous, note failures are disregarded in this mean.

As Figure(3.10) suggests assuming successful prediction using (3.8) is preferable followed by a model using station as a categorical variable. Moreover a stratified model, with respect to station, preforms worse whilst disregarding the location modeling almost gives the same result as using categoricals. Due to lack of data the loss due to late application of fungicides can not be accurately estimated.

3.9 Late blight damage

Lastly the loss in profit due to late blight and its predictability is investigated. Recall that counties are rather correlated in yield particularly considering difference in treatments. The data suggests yield does not have any significant autocorrelation nor moving average, both regarding differences between treatments and single treatments.

The damage due to late blight can be interpreted as the difference between a treated and non-treated experiment. Moreover as yield is hard to predict as its heavily influenced by weather and cultivation conditions, it will be considered known for a given treatment. However assuming the crops are treated, and there is no non-treated experiment available at the same station, this amount must be estimated in order to evaluate the difference, alternatively directly estimating the difference.

3.10 Linear regression

Trying to estimate the difference generally for all regardless station belonging of sample without repeatedly using correlated variables there is only one covariate significantly different from zero, the yield from experiment with used treatment. Moreover the variance lies around 82 tons/ha unless considering correlated covariates. Assuming everything is available for one treatment except data directly related to other treatments within the experiment, example wise consider station one and the worst treatment (T_2) then all data for the other stations may be used together with data in the county with given treatment where as data related to T_1, T_3, T_4 in C_1 will be considered unknown. This simulates a scenario where the specified station is to be evaluated where all experiments still are under progress hence the most possible amount of data.

Due to best obtained model had large variance, approximately 35 tons/ha , and application difficulties this approach was abandoned. Note regardless considering the yield or differences in yield no covariate had power to explain this variety.

Chapter 4

Discussion

Within this discussion the results, questions at issue and future possibilities will be discussed. As the data was not optimal for this experiment, due to large variation within the data set hence also this will be briefly disputed. Lastly in order to preserve the work and put it into use possible continuations and improvements will discussed.

4.1 Results

4.1.1 Data

As the no observation of data directly related to the conclusion (covariates which was removed in a later stage) was included in the imputation no large effect of this is likely to have occurred.

Degree day sum was suggested to estimate the emergence date of potatoes. As the *Phytophthora infestans* develop late blight after the emergence it is reasonable to consider this as a starting point for related covariates [8]. Emergence dates was not documented hence it was not possible to validate the method efficiency and it was assumed to be accurate enough, i.e. extra source of uncertainty. Other transformations is solely technicalities to adapt covariates in an appropriate manner. Note the analysis only covered some transformations hence there might exist some which further improves the result.

4.1.2 Regression

As previously mentioned the data varies over time, and due to lack of correlation in samples between 1983-1993 for the variables of interest these years was removed. Probably it is possible to gain more information using the removed samples if used in a smart manner.

The yield and treatments was correlated between stations in various extent. Some variety might depend on local effects, example wise one station is known to be more windy and dryer compared to the other [2]. Generally considering difference in yield, comparing treated experiments to non-treated, resulted in higher correlation, see section 3.2 Table (3.5).

Proceeding calculating profit obtained it was shown, under fixed assumptions, that it is generally preferable to use treatment programmes, see section 3.2 Figure(3.3). It also suggested that even if numerous unnecessary sprays is applied, it was profitable due to the fungicidal cost relative yield profit.

4.1.3 SIMCAST

First it is worth noting that SIMCAST associates with the optimal growth conditions given in section (1.2.1) and note the maximal temperature effect is given at the same heat and the lowest time span is the same. Moreover it is stated in [3] that southern Sweden, i.e. where the experiments was made, is more favourable for *Phytophthora infestans*.

As expected SIMCAST is rather slow moving as it denies lots of information unless the relative humidity is not high enough during a set time. Also the continuous representation, cSIMCAST derived by *Veiko Lehsten*, lacked in performance. Both had huge accuracy issues even considering a local level. Due to these even an optimized threshold gave bad results. However comparing these models it was notable that SIMCAST did not have the smooth blight unit increase that was supplied by the cSIMCAST due to its nature.

4.1.4 Cox Regression

Testing multiple covariates as single predictor Cox regression models the results did not agree with *Veiko Lehsten's* [8], i.e. the single covariate model is obtained as:

$$h(t|x) = h_0(t)e^{c(\beta(T)rH)_n^{(c)}} \tag{4.1}$$

which was second to

$$h(t|x) = h_0(t)e^{c\delta SIMCAST} \tag{4.2}$$

where c is the corresponding estimated parameters obtained in Cox regression. Proceeding, with (4.1), using manual stepwise back- and forward covariate selection[1] it was noted that the model could be further extended using year, and considerably county if available[2]. To see the effect of county, see Figure3.6 where it is shown using a stratified model.

However proceeding with the best single covariate model(4.2) resulted in the best model:

$$h(t|x) = h_0(t)e^{c_1\delta SIMCAST + c_2 PAL + c_3 Pd} \tag{4.3}$$

where $\delta SIMCAST$ is defined as in equation (3.3), PAL is benchmark value for phosphorus fertilization and Pd denotes the planting day. Note this may not be the best model but only the best model found using given data. Models obtained using elastic net combined with Cox regression was also satisfactory.

[1]

[2] All correlated covariates was intentionally avoided during this progress and individually evaluated

4.1.5 Predictions

The prediction stage was divided into two comparison stages, firstly the simple models and secondarily (also including) Cox models. Comparing SIMCAST, cSIMCAST and foliage prediction it was possible to fit thresholds such that cSIMCAST and the classical SIMCAST had similar failure rate as foliage prediction (by almost removing the thresholds completely). Due to the smoother behaviour cSIMCAST also had a more flexible level for a suitable threshold, however other issues such as bad accuracy was inherited from SIMCAST. Neither of the SIMCAST models was accurate in predictions and heavily out-preformed by foliage prediction.

The only simple method which could compare to the Cox models was foliage prediction which had the best accuracy, 1% better than Cox $\delta SIMCAST$ using parameters given in Table 3.8. However it lacks in accuracy and tend to heavily over-predict suggesting to early measures against late blight. It had similar prediction accuracy to a Cox model based on $(\beta(T)rH)_n^{(c)}$ and year as covariates.

If no hourly samples are available, it would be beneficial if its possible to consider data with respect to station, both regarding accuracy and also in number of successful attempts. However if hourly samples are available $\delta SIMCAST(3.8)$ is a generally better model as it does not need any information regarding which station that is considered. It does still outperform the others except for a stratified model which it may not be fairly compared to[3]. It is also the best performing model regarding accuracy in predictions which is the most beneficial environmental attribute.

Comparing the covariates of the best model (4.2) to those obtained by elastic net it is notable the most important seem to be related to planting and emergence dates. Then δSIMCAST follows etc.

[3]This due to the stratified model consists of three different models, hence more generous in assumptions.

4.2 Profitability

Disregarding the cost of failure and station, the most profitable method is $\delta SIMCAST$(3.8) which follows from its relative accurate predictions which is approximately 422 sek/ha more cost efficient than foliage prediction regarding fungicidal costs. Moreover its around 500 sek/ha more cost efficient than the categorical Cox model3.6.

It is notable that even though the station was not significant there is differences with respect to location and the methods may be better locally, example wise is $\delta SIMCAST$(3.8) not as efficient in C_2 as neither foliage prediction nor the categorical Cox model(3.6) where as it may not be contested in C_3 by any other model.

The best model may be beneficial, but should be further verified in field testing before complete usage. Also it would be interesting defining a strategy based on several models as a continuation.

4.2.1 Late blight damage

Estimating the late blight damage proved to be hard. Using all possible covariates and considering a few transformations did not give sufficient results. Due to the large variance in relative cost, these would not be able to use for a insurance. Including other experimental data was more successful, but did not contribute enough. Note that both the variance in late blight prediction and estimation of loss would add up further. As the damage was not successfully estimated neither would can a reasonable insurance be defined lacking of information as the large gain lies within protecting the crops and not reduced number of sprays.

4.3 Question at Issue

There are several ways to predict late blight, but it is not a trivial task due to its complex nature. SIMCAST is a historically used decision method which is based on daily mean temperature and duration of set relative humidity levels. Also biological observations may be used, example-wise as used in the experiments when foliage meets between rows.

Modelling late blight appearance using Cox regression has potential due to its simplicity and accuracy. Comparing with foliage prediction the failure rate is similar but Cox regression has better accuracy, i.e. less spraying should increase the profit if assumptions are met. Recall ten day ahead spraying is assumed to give optimal effect on yield hence earlier fungicidal application is unnecessary.

SIMCAST was lacking in optimal performance generally and was outperformed by all other methods. The contentious representation, cSIMCAST, had more potential. It was worse than foliage prediction and Cox-regression.

Defining an insurance based on Cox regression might prove possible, but then better models for the relation to yield must be constructed. Linear regression for given covariates is not sufficient as it had too large variance. However there is potential within the field if data for modelling the relation between late blight, yield and weather is available. Note that as there are several kinds of potato crops cultivated which most likely needs individual calibration etc., hence difficulties may arise along the way.

4.4 Data

As there are several conditions which are not constant, such as substances and amounts used in the cultivation (pesticides, fungicides, fertilization, and weed control) and also the *Phytophthora infestans* reproduction cycle. Beyond this there is a large natural variation. These conditions adds variety which is hard to explain using available covariates. There is more data which was not digitalized during the thesis and obtained in a late stage. The extended data contains inter alia spraying substances and further experimental details, application amount of all used substances, treatment programmes, observation dates etc. Recall the data used only had means of the best, worst and untreated experiments (which varied over time) with an estimated date of appearance. Due to all estimation stages some important knowledge might be neglected such as overall consistency and local variation.

The prices, yield per hectare and of fungicidal application, considered was also estimates and only obtained for a reference year. It is not reasonable to assume it remains constant due to facts like, a year which gives high harvest is likely to surpass the market hence lowering the prices in contrast to if all farmers gets a low harvest, the demand has decreased as new complements has become available etc. Also the sort has a significant effect on the prices which may differ over time. Bintje was the only sort used in the experiments which is considered to be susceptible to late blight.

Lastly a few notes regarding the weather related data. This part of the data set was estimated based on nearby stations and not real observations. In some extent it might be beneficial, in particular regarding relative humidity which is known to be hard to measure accurately hence a stable estimate might be preferable. Else wise it would be preferable if real observations was used. Lacking seven days ahead prognosis data both SIMCAST and Cox-regression used available data assuming relative correctness in possible prognosis, however this assumption might contribute to a lower overall variance.

4.5 Perspective

Recalling background information one may put the analysis in a larger perspective. Cultivating potatoes without fungicides reduce the yield and profit significantly. By increasing the prediction accuracy it would be possible to reduce profit loss in terms of unnecessary spraying and decreased number of failures. Using Cox regression modelin decision making would most likely be beneficial. Moreover due to its flexibility it is also possible to adjust the threshold and thereby decrease number of failures whilst increasing fungicidal application.

Due to the climate changes, it is likely that late blight will become a larger issue due to a more beneficial climate for *Phytophthora infestans* [1]. Moreover the resistance breeding proceeds slowly and treatment programmes needs constant adoption of several kinds.

All the experimental data used considered only Bintje plants which considered to be susceptible. There are other kinds with similar quality and productivity which may be preferable from this aspect.

4.6 Continuations and improvements

There is potential within the field of insurances and modeling concerning cultivation in general. As there is a vast variety which is hard to fully explain such as connection to weather and different infection risks. This thesis was only one part which considered a part of the large issue. I believe if interactions between different sources which affect the yield, such as pesticides, fungicides, general weather, soil, sorts and cultivation conditions could be included in a huge model would give a very stable model. However in order to obtain such several experiments would have to made in a large scale.

4.6.1 Mathematical considerations

If all data collected was available more covariates would be possible to include in a modeling stage and may have had a direct effect. Finding the break point in time mathematically (Monte-Carlo methods) might also add additional information, however this is only if it was not optimal already. Also further investigation of similarities and differences after the break point could allow additional information from the early experiments to the final model, example wise by defining a suitable non-linear model. For such information about the dynamics most likely also would be preferable.

There is also potential to use copulas modeling joint distribution functions and thereby enable to use more correlated covariates. Example wise a copula related to weather (relative humidity, rain, sun etc.) and *Phytophthora infestans* growth conditions could likely contribute as covariate[4] Extreme value modeling could also be possible to consider for extreme events in insurances.

Some research in the area suggest that sun (photosynthesis) affects yield positively, whereas UV exposure has a negative effect on late blight, hence this is likely to be an important covariate proceeding modeling the damage caused by *Phytophthora infestans*. Related to previous, also shade has a significant effect on the total harvest [22]. An other covariate used in other models is soil temperature which might be considered [22].

Defining a working insurance is not trivial. In order to do so a few areas need further investigation: Firstly practical aspects such as treatment programmes need to be properly investigated, costs, effect etc., enabling binding the insurance to set programmes. Also deliberate cheating would be an issue hence control methods would have to be developed ensuring the insurance holder fulfils the agreements. Lastly current modelling lacks in accuracy which preferably needs improvement.

[4]This would be similar to SIMCAST with some extensions.

Chapter 5

Summary

Potato yield is heavily negatively affected by *Phytophthora infestans* which causes late blight. These attacks has been an issue during a long time and much related research has been made such as long term forecasts and fungicidal control programs. This thesis is based on compiled experimental data, testing receptivity of Bintje, supplied by *Lars Wiik* which was collected testing control programmes during 1983-2012.

The main focus of the thesis lies within prognosis of late blight appearance which can be delayed and reduced if treated properly. Some analysis of late blight impact on the yield was included in the thesis. The data was thoroughly investigated and different methods considered. Forecast methods considered was foliage prediction, Cox regression, SIMCAST and cSIMCAST. Validation was made by cross-validation using Cox-regression where as foliage prediction is a real experiment hence assumed no validation was needed, SIMCAST and cSIMCAST preformed so badly no proper validation was considered.

Cox-regression proved to have potential as it was both simple and flexible preforming well. Regarding the considered methods the well tested foliage prediction was slightly worse than Cox. As previously mentioned SIMCAST and cSIMCAST did not have the same potential estimating the appearance date. Profit by savings from reduced number of sprays was used to evaluate the method efficiency together with accuracy and amount of failures. Using the best model savings around 422 SEK/ha can be made from reduced fungicides alone[1]. It correspond to approximately one spray less. It is assumed to be higher as the number of failures also are lower, but these numbers are not identified.

[1]This may be improved adjusting the safety threshold which was chosen such it was comparable to the currently used methods.

Using given data a model, good enough to properly model damage due to late blight, was not obtained.

Bibliography

[1] A. H. Sparks, G. A. Forbes, R. J. Hijmans, and K. A. Garrett, "Climate change may have limited effect on global risk of potato late blight," *Global change biology*, vol. 20, no. 12, pp. 3621–3631, 2014.

[2] L. Wiik, "Potato late blight and tuber yield: results from 30 years of field trials," *Potato research*, vol. 57, no. 1, pp. 77–98, 2014.

[3] H. C. Börje Olofsson, "Sortanpassad bekämpning av potatisbladmögel (phytophthora infestans)," 1994.

[4] J. B. Ristaino, "Tracking historic migrations of the irish potato famine pathogen, phytophthora infestans," *Microbes and infection*, vol. 4, no. 13, pp. 1369–1377, 2002.

[5] "SMHI wheather data." `http://www.smhi.se/kunskapsbanken/klimat/sveriges-klimat-1.6867`. Accessed: 2017-04-12.

[6] "LfL late blight information." `http://www.lfl.bayern.de/ips/blattfruechte/034444/`. Accessed: 2017-04-12.

[7] Lars Wiik, "Photo," 2015.

[8] V. Lehsten, L. Wiik, A. Hannukkala, E. Andreasson, D. Chen, T. Ou, E. Liljeroth, Å. Lankinen, and L. Grenville-Briggs, "Earlier occurrence and increased explanatory power of climate for the first incidence of potato late blight caused by phytophthora infestans in fennoscandia," *PloS one*, vol. 12, no. 5, p. e0177580, 2017.

[9] E. Vittinghoff, D. V. Glidden, S. C. Shiboski, and C. E. McCulloch, *Regression methods in biostatistics: linear, logistic, survival, and repeated measures models*. Springer Science & Business Media, 2011.

[10] N. J. Grünwald, G. R. Montes, H. L. Saldaña, O. R. Covarrubias, and W. E. Fry, "Potato late blight management in the toluca valley: Field validation of simcast modified for cultivars with high field resistance," *Plant Disease*, vol. 86, no. 10, pp. 1163–1168, 2002.

[11] S. Van Buuren, *Flexible imputation of missing data*. CRC press, 2012.

[12] Y. Wan, S. Datta, D. Conklin, and M. Kong, "Variable selection models based on multiple imputation with an application for predicting median effective dose and maximum effect," *Journal of statistical computation and simulation*, vol. 85, no. 9, pp. 1902–1916, 2015.

[13] W. Härdle and L. Simar, *Applied multivariate statistical analysis*, vol. 22007. Springer, 2007.

[14] E. Vittinghoff, D. V. Glidden, S. C. Shiboski, and C. E. McCulloch, *Regression methods in biostatistics: linear, logistic, survival, and repeated measures models*. Springer Science & Business Media, 2011.

[15] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[16] "Gödslingsråd säsongen 2017," 2017.

[17] "Midwestern Regional Climate Center gdd growing degree." `http://mrcc.isws.illinois.edu/gismaps/gddinfo.htm`. Accessed: 2017-06-01.

[18] D. Eriksson, U. Carlson-Nilsson, R. Ortíz, and E. Andreasson, "Overview and breeding strategies of table potato production in sweden and the fennoscandian region," *Potato Research*, vol. 59, no. 3, pp. 279–294, 2016.

[19] A. Haverkort, P. Boonekamp, R. Hutten, E. Jacobsen, L. Lotz, G. Kessel, R. Visser, and E. Van der Vossen, "Societal costs of late blight in potato and prospects of durable resistance through cisgenic modification," *Potato Research*, vol. 51, no. 1, pp. 47–57, 2008.

[20] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for cox's proportional hazards model via coordinate descent," *Journal of Statistical Software*, vol. 39, no. 5, pp. 1–13, 2011.

[21] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of Statistical Software*, vol. 33, no. 1, pp. 1–22, 2010.

[22] M. F. Ahmad and G. Rasul, "Prediction of soil temperature by air temperature; a case study for faisalabad.," *Pakistan Journal of Meteorology Vol*, vol. 5, no. 9, 2008.

Appendix A

SIMCAST

A.1 SIMCAST

Discrete SIMCAST based on a highly susceptible potato sort[1]. Where input is: the number of hours where the relative humidity is above 90% (time.int) and daily mean temperature. Output is blight units which may be used cumulatively for prediction of late blight appearance.

```
SIMCAST_DD <- function(time.int , temp){
    if(temp<=27 & temp>=23){
        if(time.int>=7 & time.int<=9){return(1)}
        if(time.int>=10 & time.int<=12){return(2)}
        if(time.int>=13 & time.int<=15){return(3)}
        if(time.int>=16 & time.int<=18){return(4)}
        if(time.int>=19 & time.int<=24){return(5)}
    }

    if(temp<=22 & temp>=13){
        if(time.int>=7 & time.int<=9){return(5)}
        if(time.int>=10 & time.int<=12){return(6)}
        if(time.int>=13 & time.int<=24){return(7)}
    }

    if(temp<=12 & temp>=8){
        if(time.int==7){return(1)}
        if(time.int>=8 & time.int<=9){return(2)}
        if(time.int==10){return(3)}
        if(time.int>=11 & time.int<=12){return(4)}
        if(time.int>=13 & time.int<=15){return(5)}
        if(time.int>=16 & time.int<=24){return(6)}
    }

    if(temp<=7 & temp>=3){
        if(time.int>=10 & time.int<=12){return(1)}
        if(time.int>=13 & time.int<=15){return(2)}
        if(time.int>=16 & time.int<=18){return(3)}
        if(time.int>=19 & time.int<=24){return(4)}
    }
    return(0)
}
```

[1]This model is based on SIMCAST table given in the article [8]

A.2 cSIMCAST

Combining the beta and logistic functions by multiplication the continuous variant of SIMCAST (cSIMCAST) is obtained.

A.2.1 Beta SIMCAST fit

SIMCAST fitted beta function with respect to time. Takes daily mean temperature as input, not it may be adjusted for other sampling time.

```
SIMCAST.Beta <- function(Temp) {
    minTemp = -1.9453
    optTemp = 17.1268
    maxTemp = 29.6298

    alph = log(2)/log((maxTemp-minTemp)/(optTemp-minTemp))
    bet = NA
    for(i in 1:length(Temp)){
        if( minTemp<Temp[i] & Temp[i]<maxTemp){
            bet = c(bet,
                  (2*((Temp[i]-minTemp)^alph) *
                  ((optTemp-minTemp)^alph) -
                  (Temp[i] -minTemp)^(2*alph))/
                  ((optTemp - minTemp)^(2*alph)) )
        }
    else{bet=c(bet,0)}
    }
    return(bet[2:length(bet)])
}
```

A.2.2 Logistic SIMCAST fit

A logistic (aka sigmoid function) which is fitted to SIMCAST with respect to time. Uses number of hours where the relative humidity is above 90% as input.

```
SIMCAST.Logistic <-function(Time){
    a = 7.325
    b = -0.5270
    c = 4.6006
    s = a/(1+exp(c+b*Time))
    return(s)
}
```