Stockholms universitet

Cold Winters' Influence on Insurance Claims

Stephanie Andersson

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2012:5
Matematisk statistik
Maj 2012

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Mathematical Statistics
Stockholm University
Bachelor Thesis **2012:5**
http://www.math.su.se

Cold Winters' Influence on Insurance Claims

Stephanie Andersson*

May 2012

Abstract

The winter 2010-2011 was a cold and very expensive winter for the insurance companies in the Nordic countries. The winters the past two decades have been quite warm in comparison to the winters in the past century. The question was raised whether there is a correlation between a quantified exposure to freezing temperatures and insurance claims. In that case, are the insurance companies prepared for the increase in insurance loss if the winters would be as cold as they have been the past century; how large can the losses be expected to be in the three regions of Sweden (Norrland, Svealand and Götaland) within each sector (Private, Agricultural and Commercial)? Data was gathered from a Swedish insurance company regarding the insurance claims between 1985-2011. The temperatures were gathered from SMHI (Sveriges Meterologiska och Hydrologiska Institution) and ECAD (European Climate Assessment Dataset project). The temperatures were used to construct temperature indexes for the winter period. We tested two different indexes based on the daily minimum and daily mean temperatures, showing that the minimum temperature index was generally not as strongly correlated with the insurance claims as was the mean temperature index, why the mean temperature index was used for the remainder of the analysis. In order to investigate the relationship between the temperature index and insurance claims, a correlation analysis and then a simple linear regression analysis were preformed demonstrating a strong correlation between mean temperature index and insurance claims. Prediction analysis was made regarding the years between 1930-1985. When we examined the results, we could see that a winter as cold as the winter 1941-42 would generate big insurance claims in comparison with an average winter from the past two decades. However, the prediction from this model is very uncertain, which is demonstrated with a wide 95 % prediction interval. The report ends with a discussion on the performed analysis and results, where limitations and other sources of insecurities are mentioned.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: stephanieanderss@hotmail.com . Supervisor: Jan-Olov Persson.

Abstract

The winter 2010-2011 was a cold and very expensive winter for the insurance companies in the Nordic countries. The winters the past two decades have been quite warm in comparison to the winters in the past century. The question was raised whether there is a correlation between a quantified exposure to freezing temperatures and insurance claims. In that case, are the insurance companies prepared for the increase in insurance loss if the winters would be as cold as they have been the past century; how large can the losses be expected to be in the three regions of Sweden (Norrland, Svealand and Götaland) within each sector (Private, Agricultural and Commercial)? Data was gathered from a Swedish insurance company regarding the insurance claims between 1985-2011. The temperatures were gathered from SMHI (Sveriges Meterologiska och Hydrologiska Institution) and ECA&D (European Climate Assessment Dataset project). The temperatures were used to construct temperature indexes for the winter period. We tested two different indexes based on the daily minimum and daily mean temperatures, showing that the minimum temperature index was generally not as strongly correlated with the insurance claims as was the mean temperature index, why the mean temperature index was used for the remainder of the analysis. In order to investigate the relationship between the temperature index and insurance claims, a correlation analysis and then a simple linear regression analysis were preformed demonstrating a strong correlation between mean temperature index and insurance claims. Prediction analysis was made regarding the years between 1930-1985. When we examined the results, we could see that a winter as cold as the winter 1941-42 would generate big insurance claims in comparison with an average winter from the past two decades. However, the prediction from this model is very uncertain, which is demonstrated with a wide 95%prediction interval. The report ends with a discussion on the performed analysis and results, where limitations and other sources of insecurities are mentioned.

Special Thanks To

I am truly grateful and appreciative of all the colleagues, friends and family who have contributed to this thesis.

I would like to direct my gratitude towards my tutor by the Mathematical Institution at Stockholm University, Jan-Olov Persson for all the guidance and support he has given me. Thank you for being an excellent supervisor and for helping me through thick and thin.

Thank you Olof Fält and Florent Scarabin for accepting me at Guy Carpenter and for always having time to answer questions, helping to solve problems and for your invaluable help, personal commitment and patience.

Thank you Mikael Finder and David Kohlberg, my big supporters, for always being there for me and for your never-ending support.

Contents

1 Introduction

Insurance is a form of risk management used to avoid the risk of a conditional, uncertain loss. An insurer is a company selling the insurance; the insured, or policyholder, is the person, company or entity buying the insurance policy. Risk management, the practice of assessing and controlling risk, has evolved as a field of study and practice. Property insurance provides protection against risks to property, such as fire, theft or weather damage.

Reinsurance is a form of risk management that one insurance company sells to another insurance company. The reinsurance main purpose is to protect the insurance company from large insurance claims due to catastrophes[1]. Guy Carpenter is a global risk and reinsurance specialist. They provide insurance companies with in-depth analysis in a number of areas, from market conditions, to catastrophe analysis, to environmental issues[2]. They were approached by a Swedish insurance company and asked to analyze the relationship of cold winters and the insurance company's insurance claims. The reason for this was to evaluate the insurance company's reinsurance policies, since the past two years had resulted in many large claims regarding property damages due to freezing. Is the insurance company's current reinsurance program sufficient to cover for potential insurance claims due to freezing damages?

Due to a signed agreement with the insurance company, no data regarding the sizes of the insurance company or the insurance claims may be disclaimed, why some plots, numbers and names have been omitted from this thesis.

[1] Gustafsson B, Återförsäkring, Tierps Tryckeri AB/ Partener Print, 2000 s. 7-8,10
[2] www.guycarp.com

2 Background

The past two winters have been excessively expensive for the insurance company. The reason for this is large claims resulting from broken water pipes due to cold winters. Therefore the insurance company has showed an interest in demonstrating a correlation between cold winters and large insurance claims.

The insurance company provided for their data on daily insurance claims on property insurance from three different sectors: Private, Agricultural and Commercial. The data consists of their daily records on property insurance claims from 1979-2011. Due to a signed confidentiality agreement the name of the insurance company is not revealed in this thesis.

The World Meteorological Organization (WMO) is a specialized agency of the United Nations. WMO promotes cooperation in the establishment of networks for making meteorological, climatological, hydrological and geophysical observations, as well as the exchange, processing and standardization of related data, and assists technology transfer, training and research[3].Two organizations working with WMO that registers Swedish weather conditions are:

- SMHI (Sveriges Meterologiska och Hydrologiska Institution). SMHI is a government agency under the Ministry of the Environment. SMHI's mission is to manage and develop information on weather, water and climate that provides knowledge and advanced decision-making data for public services, the Private sector and the general public. SMHI receives daily data from 52 different temperature stations in Sweden.[4]

- The ECA&D (European Climate Assessment Dataset project) forms the backbone of the climate data node in the pilot Regional Climate Centre (RCC) for WMO Region VI (Europe and the Middle East) since 2010. It receives data from 58 participants for 62 countries and the ECA&D dataset contains 31058 series of observations for 12 elements at 6596 meteorological stations throughout Europe and the Mediterranean. 38% of these series are public, which means downloadable from their website for non-commercial research. Participation to ECA&D is open to anyone maintaining daily station data. [5]

[3]http://www.wmo.int/pages/about/index_en.html
[4]http://www.smhi.se/en/about-smhi
[5]http://eca.knmi.nl

3 Aims and Methods

The purpose of this thesis is to examine if and how the freezing temperatures during wintertime affect the insurance claims on property during the same period. Is there a stronger relationship in some areas? How big would the claims be today if the winters would be as cold as the coldest period of this century?

The insurance company registers their daily insurance claims geographically all over Sweden. The data consists of the insurance claims divided into three sectors – Private, Agricultural or Commercial – the claims come from. The claims have been divided into amount paid and amount in reserve.

Two factors are involved in damages caused by freezing.

1. Temperature (the number of degrees Celsius below freezing point – 0°C)
2. Number of days with a temperature below freezing.

Therefore one temperature index that is derived from the daily mean temperatures and one temperature index that is derived from the daily minimum temperatures from the winter period was created with the aim to address both of these factors. The temperature index is more deeply described under the chapter "Data Description".

To achieve our aims, the analysis will be divided into three steps:

1. Scatter plots to examine the relationship between the mean temperature indexes and insurance claims. Is the mean temperature index correlated with the insurance claims? Is there a stronger relationship in any area of Sweden? We then preform a simple linear regression analysis to construct a model and to examine the relationship between the insurance claims and the temperature index.
2. Comparison between a simple linear regression analysis of the explanatory variable represented by the minimum temperature index and the mean temperature index. Which one of minimum and mean is better to use to predict insurance claims? We will preform a multiple linear regression with both minimum and mean temperature index to see if this method improves our model.

3. Prediction using the most appropriate temperature index above in order to examine lost insurance claim history data between 1930 and 1985 (the insurance company started registering accurate daily insurance claims 1985). How big would the claims be today if the winters would be as cold as the coldest period this century, using whichever of minimum- and mean temperature index that proves more reliable in step 2?

To analyze if the temperature index has any influence on insurance claims we have focused on the correlation coefficient and simple and multiple linear regression. The insurance claims have been adjusted with regards to inflation and portfolio development. Predictions have been made using the prediction method and prediction intervals. As diagnostic methods of model criticism we have calculated the coefficient of determination, level of significance, confidence intervals, prediction intervals, studied potential autocorrelation and made a residual analysis.

All data provided is purely numerical.

Calculations have been made with SAS, Microsoft Excel 2007.

4 Methods and Terms Used

The terms used in this thesis are taken from the compendium "Lineära Statistiska Modeller" by Rolf Sundberg, "Statistik – metoder och tillämpningar" by Gunnar G. Løvås and "Stokastiska Metoder" by Sven Erick Alm and Tom Britton.

4.1 Correlation

The correlation coefficient is a measure of the linear correlation between two variables and is designated with ρ. The formula for sample estimation:

$$\hat{\rho} = r_{xy} = \frac{c_{xy}}{s_x s_y},$$

where $c_{xy} = \frac{1}{(n-1)} \sum (x_i - \overline{x})(y_i - \overline{y})$

$$s_x = \sqrt{\frac{1}{(n-1)} \sum_{i=1}^{n} (x_i - \overline{x})^2}$$

$$s_y = \sqrt{\frac{1}{(n-1)} \sum_{i=1}^{n} (y_i - \overline{y})^2}$$

n is the number of observations. The value of the correlation coefficient is between -1 and $+1$, where $r_{xy} = 1$ indicates a positive correlation, whereas $r_{xy} = -1$ indicates a negative correlation. $r_{xy} = 0$ suggests that there is no correlation.

4.2 Regression Analysis

The regression, as opposed to the correlation analysis does not only show if there is a correlation, but rather what the relationship is. It studies relationships between a response variable and one or more independent explanatory variables. The aim with the regression analysis is to find a linear equation that fits the observations value and through the method of least squares, minimize the sum of the squared residuals. It is called simple linear regression when using only one explanatory variable, while multiple linear regression uses more than one explanatory variable. Both are used in this thesis.

4.2.1 Simple Linear Regression

Simple linear regression is the least squares estimator of a linear regression model with only one explanation variable.

The definition of the simple linear regression model can be written:

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

where y_i is the response variable and represents a value y for the observation i. x_i is the explanatory variable, α and β are parameters, ε_i is the random error variable and i = 1,2...,n where n is the number of observations. The model requires that the ε_i are independent among themselves and normally distributed with N(0,σ^2). α is called intercept and determines the line's intersection of the y-axis. β is the gradient coefficient and determines the slope of the line.

The simple regression has an expectation equation:

$$E(y_i) = \alpha + \beta x_i,$$

where the parameters have same meaning as above.

The formula for the estimated parameters $\hat{\alpha}$ and $\hat{\beta}$:

$$\hat{\beta} = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sum_i (x_i - \overline{x})^2}$$

$$\hat{\alpha} = \overline{y} - \hat{\beta}\overline{x}$$

4.2.2 Multiple Linear Regression

When making a linear regression on a dataset, where there are more explanatory variables able to influence the response variable, you have to use the multiple linear regression analysis. The formula for multiple linear regression is:

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_m x_{mi} + \varepsilon_i$$

where y_i is the response variable, $x_{1i}, x_{2i}, \ldots, x_{mi}$ are the explanatory variables and $\alpha, \beta_1, \ldots, \beta_m$ are parameters, ε_i is the random error variable and $i = 1, 2, \ldots, n$ where n is the number of observations and m is the number of variables. The model requires the random variables to be independent among themselves and normally distributed N(0,σ^2).

4.2.3 Residual

The residuals are used to see if the regression model fits the data. It is important to discover dependent and systematic errors. The equation for the residual can be written:

$$\hat{\varepsilon}_i = y_i - \hat{y}_i$$

4.2.4 Coefficient of Determination R^2

The coefficient of determination R^2 is a measurement of how much of the variability in a data set is "explained" by the statistical model. The proportion of explained variation is given by:

$$R^2 = \frac{KVS_{Regression}}{KVS_{Total}} = 1 - \frac{KVS_{Residual}}{KVS_{Total}} = r_{xy}^2$$

$KVS_{Total} = \sum(y_i - \overline{y})^2$

$KVS_{Regression} = \sum(\hat{y}_i - \overline{y})^2$

$KVS_{Residual} = \sum(y_i - \hat{y}_i)^2$

The coefficient of determination is a number between 0 and 1. The higher the number, the stronger the linear correlation and the more variability is explained by the statistical model.

4.3 Heteroscedasticity

Data can be considered to be heteroscedastic if the variance for one of the variables is not constant for every value of the other variable. In order to make a linear regression, the random variables need to be independent and normally distributed with a constant variance, why heteroscedasticity would hinder the method.[6]

4.4 Transformation

In the cases when the response variable y is not linear in the parameters α and β, a suitable transformation of the y-variable can be helpful to construct a linear model from a non-linear model. One common transformation is the log-transformation. Log-transformation can be helpful to minimize the heteroscedasticity or to get distorted data more normally distributed.

[6] Vejde O. et Leander E, Ordbok i statistik, Olle Vejde Förlag, 2000 s. 107 and Andersson P. & Tyrcha J. *Kompendium* s. 85-86

Some log-transformed models can be:

Transformation of y only:

$$log(y) = \alpha + \beta x + \varepsilon_i$$

$$y = e^{\alpha} \cdot e^{\beta x} \cdot e^{\varepsilon_i}$$

Transformation of y and x:

$$log(y) = \alpha + \beta log(x) + \varepsilon_i$$

$$y = e^{\alpha} \cdot x^{\beta} \cdot e^{\varepsilon_i}$$

where ε_i is the error with N(0, σ^2).

4.5 Shapiro-Wilk Statistic

The Shapiro-Wilk test for normality tests the null hypothesis that a sample came from a normally distributed population. The test rejects the hypothesis of normality when the p-value is low, which would indicate a non-normal distribution.[7]

4.6 Autocorrelation

Autocorrelation is the phenomenon where a relationship exists between the errors separated by time in a regression analysis. Andersson and Tyrcha define autocorrelation as a situation where "...the covariance between the disturbances does not depend on calender time but only on the time difference between the observations" when the criterion that "time-series data in which we interpret the number of the observation as indicating the time at which the observation was made" is met. [8]

4.7 Durbin-Watson Statistic

The Durbin-Watson (DW) statistic is a test used to detect the presence of autocorrelation.

The formula used for the DW statistic is written:

[7] http://www.jmp.com/support/faq/jmp2085.shtml

[8] Andersson P. & Tyrcha J. *Kompendium* s. 62, 87-89

$$d = \frac{\sum_{t=2}^{T}(\varepsilon_t - \varepsilon_{(t-1)})}{\sum_{t=1}^{T}\varepsilon_t^2}$$

where T is the number of observations. ε_t and $\varepsilon_{(t-1)}$ is the residual with the observation at time t. The interval for d will be $0 < d < 4$, where d indicates the autocorrelation. The value of d is close to 2 if the errors are uncorrelated. There is evidence of a positive autocorrelation if d is substantially lower than 2. From the test we get the 1^{st} order of autocorrelation, which will have the interval $-1 < autocorrelation < 1$. The test will also give us a p-value for positive and negative autocorrelation.[9]

4.8 Confidence Interval

The formulas given in the section ”simple linear regression” allow one to calculate the point estimates of α and β, that is, the coefficients of the regression line for the given set of data. However, those formulas do not tell us how precise the estimates are. That is, how much the estimators $\hat{\alpha}$ and $\hat{\beta}$ can deviate from the "true" values of α and β. The latter question is answered by the confidence intervals for the regression coefficients.

In order to create confidence intervals for a regression line, one uses the standard error of the regression and the t-distribution under the assumption that the standard errors are normally distributed.

Confidence interval for y_0 :

$$\hat{\alpha} + \hat{\beta}(x_0) \pm t_{p/2}(n-2)\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0-\overline{x})^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}}$$

With a set confidence level (1-p) at for example 95%, we are 95% certain that the interval covers the true value of y.

The confidence level for $\hat{\beta}$:

$$\hat{\beta} \pm t_{p/2}(n-2)\hat{\sigma}/\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}$$

where $\hat{\alpha}$, $\hat{\beta}$ are the same as above
t = t-disruption with (n-2) degrees of freedom
$\hat{\sigma}$ =the estimated standard deviation
p = significance level
n = the number of observations.

[9]Andersson P. & Tyrcha J. *Kompendium* s. 65

4.9 Prediction

Prediction is a method where values are estimated from the dataset, which means that you for a new observation x_0 can predict a response value y_0. In short, the method uses known values in order to estimate unknown values. In this thesis the prediction will be associated with regression models.

The predicted value $\hat{y_0}$ is an estimation by E(y)= $\hat{\mu} = \hat{\alpha}+\hat{\beta}x_0$. Hence the formula for simple linear regression becomes:

$$\hat{y_0} = \hat{\alpha} + \hat{\beta}x_0.$$

where $\hat{y}$ is an estimated response variable to y, $\hat{y}$, $\hat{\alpha}$ and $\hat{\beta}$ are estimate parameters and x is the explanatory variable.

The prediction error is given by:

$$\hat{\varepsilon}_i = y_i - \hat{y}_i$$

The prediction interval is an estimate of an interval within which new observations will fall, with a certain probability, given what has already been observed. If the prediction interval is wide, the predictions will not be reliable. A wide prediction interval is linked with a small R^2, which means that the regression line only will explain a small part of the variation in the data material. The prediction interval is given by:

$$\hat{\alpha} + \hat{\beta}(x_0) \pm t_{p/2}(n-2)\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0-\overline{x})^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}}$$

where $\hat{\alpha}$, $\hat{\beta}$ are the same as above
t = t-disruption with (n-2) degrees of freedom
$\hat{\sigma}$ =the estimate standard deviation
p = significance level
n = the number of observations.

5 Data Description

In this chapter, the temperature data and the insurance data will be treated separately as preparation for the statistical analysis. No match or comparison will be made until the chapter "Statistical Analysis".

5.1 Insurance Claims

The insurance company provided for their data on daily insurance claims. The data was for property insurance from three different sectors: Private, Agricultural and Commercial. The data consists of their daily records on property insurance claims from 1979-2011. Due to the signed confidentiality agreement, all sums indicating the sizes of the insurance claims have been omitted.

The dates from the insurance claims in the dataset are the dates when the insurance loss was filed. This does not necessarily mean that the damage occurred at the same date. The claims are made due to damage on property. We have no exact data that in detail specifies the cause of the claims.

The winter period for insurance claims was set from November 1^{st} to June 30^{th} and the daily claims were summed, so that we got a total for the entire period. The time period was chosen with help of Guy Carpenter. They have the experience that damages that occurred in the wintertime on summerhouses might not be discovered until summer. The period was also chosen to take the probable delay of the claim into account.

In the dataset, claims have been subdivided into "payments" and "reserve". The payments account for the money paid to the claimant and the reserve account for the money that the claimants have yet to receive. These two posts have been added in the calculations, since both are a measure of the degree of the damage claimed. The inflation and the portfolio development for the Private sector, Agricultural sector and the Commercial sector have been multiplied with the insurance claims respectively. The portfolio development for each sector is measured with the number of policies. The inflation numbers have been provided by SCB and the development of number of policies have been provided by the insurance company.

Initially, the relationship between the temperature index and the claims from each geographical area was analyzed with scatter plots, but no relationship was seen. We also noted that the claims from each geographical area were not large enough to show any relationship with the temperature data, why the areas had to be merged into larger regions.

The areas were grouped into three major regions of Sweden: Norrland, Svealand and Götaland as seen in *Image 1*. 18.59% of the total claims provided by the insurance company were from Norrland, 22.13% from Svealand and 59.28% from Götaland.

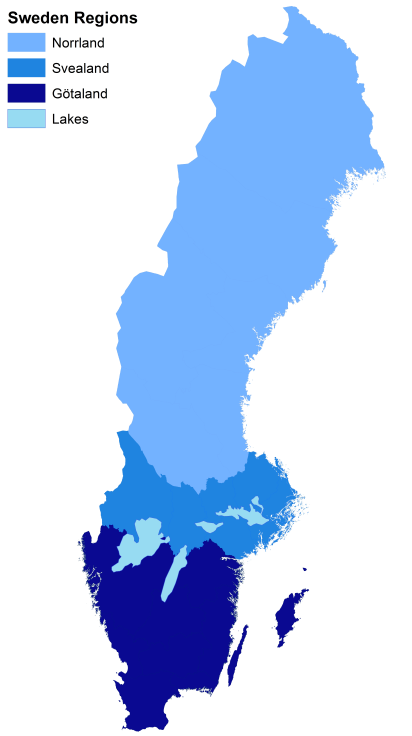


Image 1: Map of Sweden with Regions

The earliest claim registered was from 1979. From most areas in Svealand and Götaland, the insurance claims were not registered until 1980. From most areas in Norrland, we had not received any insurance data before 1992, why this region was studied regarding the years 1992-2011. Most of the other areas did not have sufficient insurance data before 1985, why we chose to disregard from all claims registered between 1979-1985 and Svealand and Götaland were studied regarding the years 1985-2011.

In Sweden as a total, the distribution of insurance claims between the three sectors was as follows: Private 46%, Agriculture 12% and Commercial 42% year 1979-2011.

5.2 Temperature Data

The daily temperatures have been provided by SMHI and ECA&D. Both organizations have data on three alternative daily temperature measurements: minimum, maximum and mean.

Both organizations display their temperatures in °C. The formula used for calculating the daily mean temperature has been weighted with regard to minimum and maximum temperatures of the day and the temperatures at 7am, 1pm and 7pm respectively.

Due to the thesis' financial limitations SMHI only provided us the daily mean temperatures between 1961-2011 from all of their 52 weather stations throughout Sweden. ECA&D had public data on daily minimum, maximum and mean temperatures in 63 Swedish cities and 6 cities in Denmark between the years 1875 and 2011.

In the areas and time-intervals where data was available from both organizations they proved to be identical. Primarily the SMHI-stations will be used, since they had more stations and more data per station. ECA&D was used because primarily for the minimum and because in some stations we had access to data starting from 1875 as opposed to SMHI that had data starting from 1961, why we needed the ECA&D-data in this thesis in order to predict the insurance claims before 1961.

In some areas, several temperature stations were represented, while the insurance company only had one source of data. Hence, we chose – if possible – the temperature station in the city with the biggest population, since this would be station best represented amongst the insurance claims. Some temperature stations were missing daily temperatures, why we for some of the areas had to choose a station in a city with a smaller population. For some areas we could not find a suitable temperature station at all.

The stations used were grouped into the same three regions as before: Norrland, Svealand and Götaland, demonstrated in *Table 1*.

Region	*SMHI Station*	*ECA&D Station*
Norrland	Luleå, Bjuröklubb, Sveg, Sundsvall, Borlänge	Luleå, Sundsvall
Svealand	Karlstad, Gustavsfors, Uppsala, Stockholm	Stockholm, Karlstads Flygplats
Götaland	Gothenburg, Jönköping, Målilla, Bredåker, Kalmar, Malmö, Torup, Skåne, Visby	Gothenburg, Jönköping, Visby, Bredåker

Table 1: The three regions of Sweden and which SMHI- and ECA&D temperature stations were included

The ECA&D stations were selected due to them being identical to the SMHI-stations.

5.3 Temperature Index

As mentioned in the section "Aims and Method" we construct an index for each winter season over the temperature that "incorporates" the number of days with freezing temperatures as well as the temperature itself.

The temperature index was created with the help of Guy Carpenter due to their large experience in the field. The days with temperatures above freezing were removed since we wanted to investigate how temperatures below freezing affected the insurance claims. Then the daily mean temperatures (SMHI) below 0°C for the period were summed and multiplied with -1, so that we got a positive "temperature index" for the negative temperatures for the entire period. When summed this way, a large number represents either a longer period with temperatures below freezing or colder daily mean temperatures. The reason for this is to get the most intuitive relationship between coldness and the time below freezing.

The same thing was done with the ECA&D daily minimal and mean temperature.

The "winter season" for each year was set from November 1^{st} to March 31^{st}.

6 Statistical Analysis

As stated before, we noted that the claims from geographical areas were not large enough to show any relationship with the temperature data, why the areas had to be merged into larger regions.

When we merge the insurance areas, we have to create a temperature index for the entire region. In order to create such an index, the temperature data needs to be weighted.

6.1 Weighting the Temperature Index

In order to create a temperature index for the three larger regions, Norrland, Svealand och Götaland, we wanted to aggregate the data through weighting the temperatures with regards to the portfolio development for each of the insurance company's areas. Due to limitations in the data supplied by the insurance company, we had to abandon this idea and had to weight the temperatures with consideration to the total insurance claims of the area, since some areas had much bigger insurance claims than others. Consequently the areas with larger insurance claims would have a higher level of correlation with the region's temperature index. The formula for the weighted temperature shown below was implemented for each year:

$$T_i = \frac{q_i}{q} t_i$$

where T_i= weighted temperature for the winter season for the local temperature station i.
q_i = total claims from the geographical area i between 1985-2011
$q = \sum_i q_i$
t_i = temperature index for the winter season from the local temperature station
i = geographic areas

After we had weighted the temperatures, the data was aggregated to form a region temperature index.

$$x_k = \sum_i^N T_i$$

where x = temperature index for the region for the selected winter period
k = region
N = number of temperature stations in the region

6.2 Analysis Strategy

As mentioned in “Aims and Methods”, the analysis will be divided into three steps:

1. Scatter plots and simple linear regression analysis of the SMHI-dataset in order to **examine the relationship** between freezing temperatures and insurance claims. The reason for using SMHI in this section is because they have more stations.

2. **Comparison** between a simple linear regression analysis of the insurance claims and the minimal- and of the mean temperatures. Which one of **minimum and mean** is better to use to predict insurance claims? Here we use ECA&D, since we had access to the organization’s minimum temperature data. We will also test a multiple linear regression with both min and mean to see if this method improves our model.

3. **Prediction** using the most appropriate temperature index above in order to examine lost insurance claim history data between 1930 and 1985. How big would the claims be today if the winters would be as cold as the coldest period this century using whichever of minimum and mean proves more reliable in step 2?

6.3 Examining the Relationship

6.3.1 Scatter Plots

Scatter plots, *Figure 1-3,* were constructed to visualize the relationship between the insurance claims reported in each region and the mean temperature index from SMHI. This was done for the three different sectors: Private, Agricultural and Commercial. Underneath each set of figures, the correlation coefficients for each sector and region is presented.

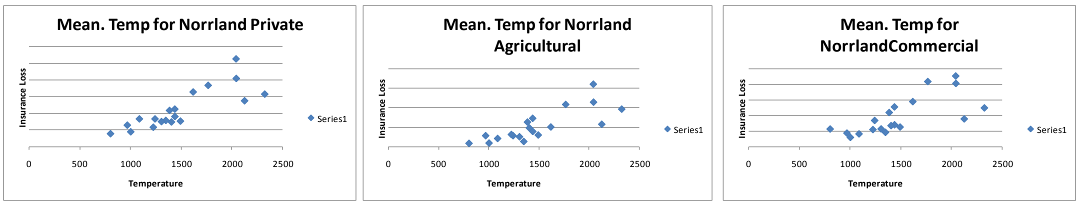


Figure 1: Scatter plots of Insurance claims in relation to mean temperature index from SMHI for Norrland Correlation coefficient– Private: r=0.83, Agricultural: r=0.80 Commercial: r=0.72

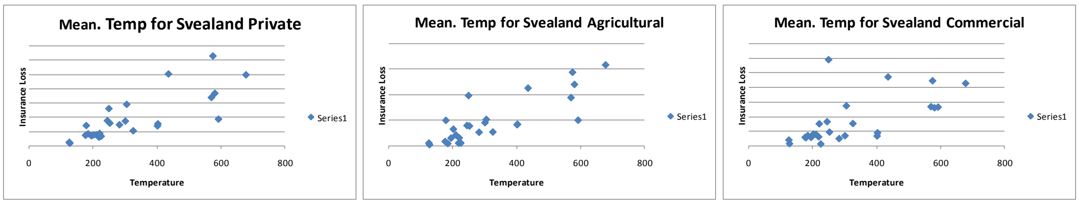


Figure 2: Scatter plots of Insurance claims in relation to mean temperature index from SMHI for Svealand Correlation coefficient– Private: r=0.80 , Agricultural: r=0.82 Commercial: r=0. 62

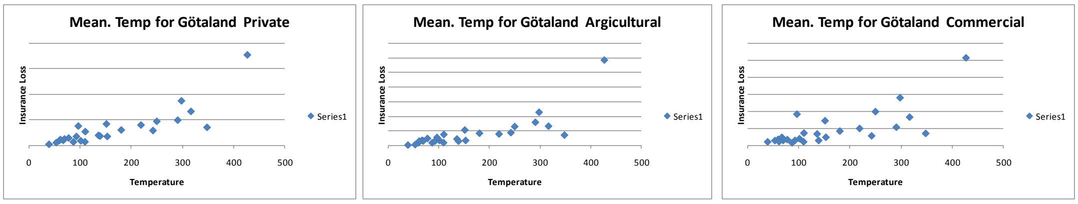


Figure 3: Scatter plots of Insurance claims in relation to mean temperature index from SMHI for Götaland Correlation coefficient – Private: r=0.82, Agricultural: r=0.77 Commercial: r= 0.74

Consistently throughout the scatter plots, a positive correlation is seen between the insurance claims and increasing temperature index with the correlation coefficients between 0.62-0.83. However, the spread increases with an increase in temperature index implying heteroscedacity. Hence, the conclusion is drawn that the linear regression analysis should be made on log-transformed variables. The reason for this is to see if we can minimize heteroscedacity. Log-transformation is suitable for data where you can see that the residuals get bigger for bigger values of the dependent variable as seen in the scatter plots above.

Log-transformation is firstly used on only the response variable, thereafter on the response- and explanatory variable together. The linear relationship improves when we log-transform the response variable, but an even stronger linear relationship is seen when both variables are log-transformed.

Scatter plots, *Figure 4-6*, are created over the relationship between the log-transformed variables: log(Private), log(Agricultural), log(Commercial), versus log(temperature index). Within these plots we include a regression line that is adapted to the observations according to the least square method (MK). The regression line simplifies visualizing the spread. Underneath each set of figures, the correlation coefficients for each sector and region is

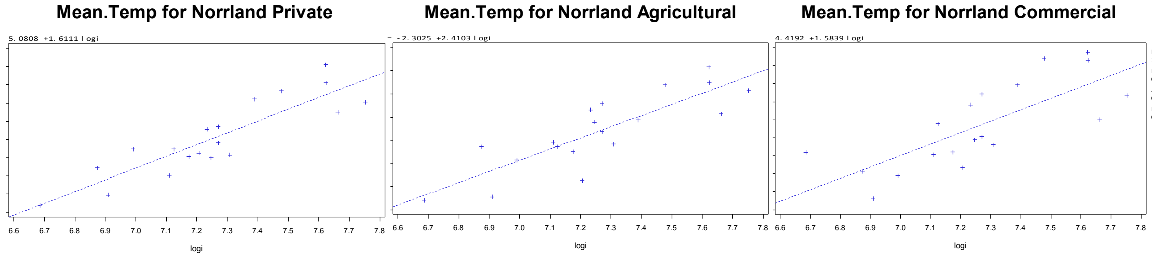


Figure 4: Scatter plots of log(Insurance claims) in relation to log(mean temperature index) with regressions line for Norrland. Correlation coefficient– Private: r=0.88, Agricultural: r= 0.84, Commercial: r= 0.76

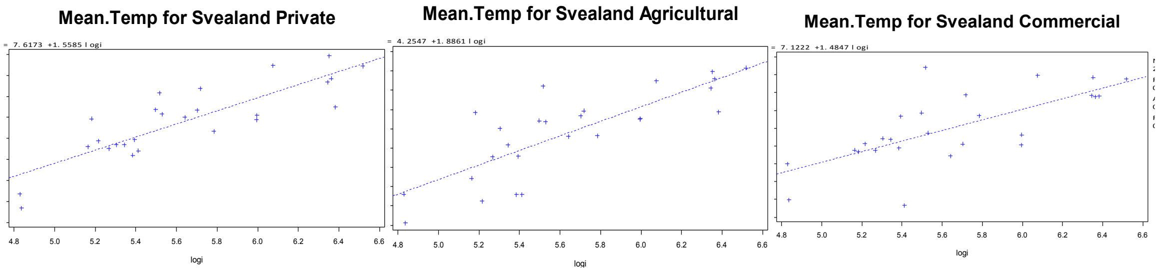


Figure 5: Scatter plots of log(Insurance claims) in relation to log(mean temperature index) with regression line for Svealand. Correlation coefficient – Private : r=0.85, Agricultural: r=0.80, Commercial: r=0.71

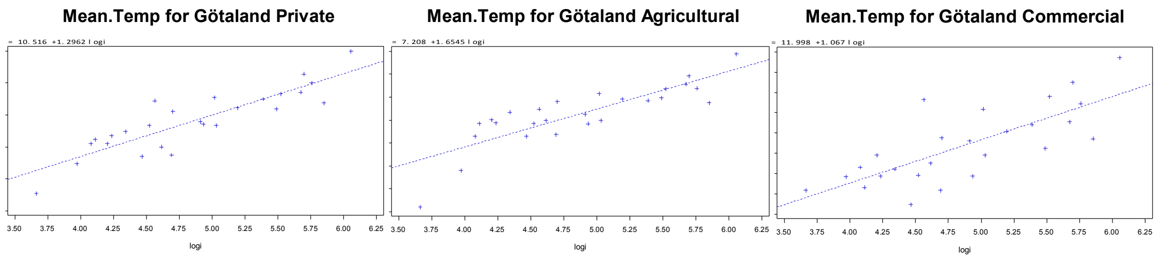


Figure 6: Scatter plots of log(Insurance claims) in relation to log(mean temperature index) with regression line for Götaland. Correlation coefficient – Private: r=0.88, Agricultural: r=0.83, Commercial: r=0.76

In the figures above a more linear relationship seems to have been established through the log transformation and the heteroscedacity seems to have decreased in all three regions. We also see that the correlation has increased with the log-transformation.

We hereby conclude that the log-transformed variables approximately have met the criteria necessary for a linear regression analysis.

6.3.2 Linear Regression

Through linear regression we can analyze the material further and look at residual plots to investigate whether the residuals are independent and normally distributed. The purpose is to detect if our response log(insurance claims) are linearly dependent or described by our explanatory variable log(temperature index).

The linear regression will consist of a simple linear regression since we only have one explanatory variable, which will be named “index”=x below. Insurance loss for all sectors, Private, Agricultural and Commercial will be named y.

We start by presuming the following formula:

$$y_i = e \cdot x^{\beta} \cdot e^{\varepsilon_i}$$

where α is the intercept, β is slope and ε_i is the error term.

By using log-transformation on the variables above, we get a linear formula as follows:

$$log(y_i) = \hat{\alpha} + \hat{\beta} log(x_i) + \varepsilon_i$$

for every sector respectively, where $\hat{\alpha}$ and $\hat{\beta}$ are estimates of α and β and i is the observation and ε_i is the error term.

Before we can study the model further, we have to control if the residuals have any systematic or dependent deviance and if the residuals are normally distributed. In order to do this we construct plots with the residuals over the predicted values, QQ-plots and preform a Shapiro-Wilk W-test.

When we have created graphs with the residuals on the y-axis and the prediction on the x-axis, we find no apparent system within the residuals, why we conclude that they are random and independent. In the QQ-plot, we noticed that the residuals were close to the normal distribution line. All plots can be found in Appendix I. For some of the sectors however, we see that the residuals are further away from the normal distribution line, indicating that the residuals are not normally distributed, why we have to test this further. Through a Shapiro-Wilk W-test we can test the null hypothesis that a sample came from a normally distributed population. The test rejects the hypothesis of normality when the p-value is low. From the Shapiro-Wilk W-test we get high p-values. Hence we cannot reject the hypothesis of normality. In the *Figure 7* below, controlling the Agricultural sector in Norrland, we get a p-value of 0,9958 indicating that the data is normally distributed.

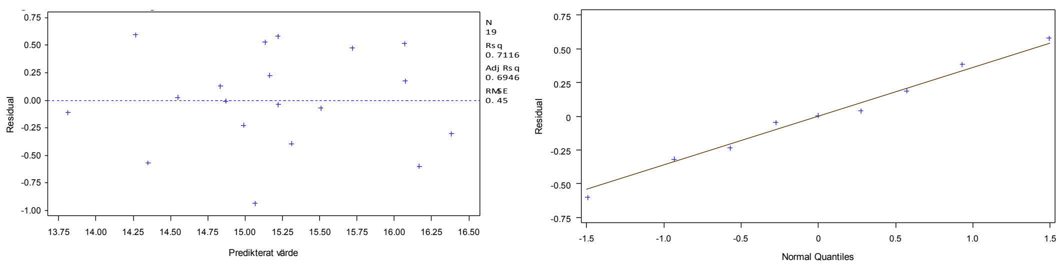


Figure 7: Norrland and Agricultural

Left: Residual over prediction

Right: A normal QQ-plot for the residuals

We also create plots showing the residuals over year to examine any potential autocorrelation. The *Figure 8* for the Agricultural sector in Norrland is demonstrated below. The other plots can be found in Appendix I.

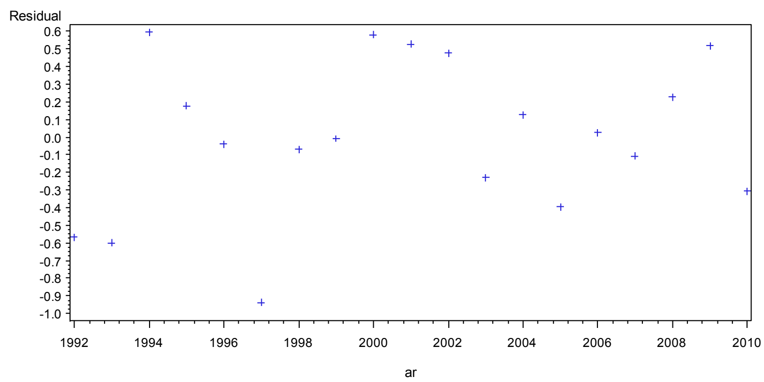


Figure 8: Norrland and Agriculture – Residual over year

In the graph, we anticipated to see a tendency towards autocorrelation, which would imply that the temperature one year affects the insurance claims the year after. In these plots we suspect a tendency towards autocorrelation, why this must be investigated further. Autocorrelation would indicate that the residuals are not random and independent. Therefore we preform the Durbin-Watson test and see that d for all regions and sectors ranges between 1,3 and 2,3 indicating non-autocorrelation. High p-values indicate a low significance for the autocorrelation and autocorrelation coefficient is 0.09.

In conclusion there appears to be a relationship between insurance claims and temperature index without autocorrelation, why we now can proceed with the analysis.

To see what influence the temperature index has on the insurance claims, we will, by using SAS, examine the confidence intervals for $\hat{\beta}$ and prediction intervals for new observations of y_i as well as R^2 and level of significance.

From SAS we get point estimations for α and β. Through a 95% confidence interval we get the discrepancy from the true value of β. We also get a p-value for the hypothesis $\hat{\beta}$=0.

To get a better overview if the log-transformation gives a linear explanation, we create a table for each sector in all three regions shown in *Tables 2-4*:

Norrland Mean Temp.	$\hat{\beta}$-value	*95% Conf. for* $\hat{\beta}$	R^2	*p-value*
Private	1.61	(1.17, 2.05)	0.78	<0.0001
Agricultural	2.14	(1.63, 3.19)	0.71	<0.0001
Commercial	1.58	(0.81, 2.70)	0.58	0.0001

Table 2. Linear regression analysis with log(y_i) and log(x_i) from Norrland the with R^2-values, p-values and β-values and their respective confidence intervals

Svealand Mean Temp.	$\hat{\beta}$-value	*95% Conf. for* $\hat{\beta}$	R^2	*p-value*
Private	1.56	(1.15, 1.97)	0.72	<0.0001
Agricultural	1.89	(1.30, 2.47)	0.65	<0.0001
Commercial	1.48	(0.86, 2.12)	0.50	<0.0001

Table 3. Linear regression analysis with log(y_i) and log(x_i) from Svealand the with R^2-values, p-values and β-values and their respective confidence intervals

Götaland Mean Temp.	$\hat{\beta}$-value	*95% Conf. for* $\hat{\beta}$	R^2	*p-value*
Private	1.30	(1.00, 1.60)	0.77	<0.0001
Agricultural	1.65	(1.19, 2.12)	0.70	<0.0001
Commercial	1.07	(0.68, 1.45)	0.58	<0.0001

Table 4. Linear regression analysis with $log(y_i)$ and $log(x_i)$ from Götaland the with R^2-values, p-values and β-values and their respective confidence intervals

In all regions there is a significant relationship between insurance claims and temperature index at a 5% level. This is seen through a p-value less than 0.0001.

The reason for the interest in the $\hat{\beta}$ values is demonstrated with the equations below.

We use the formula with the log-transformed variables below:

$$y = e^{\hat{\alpha}} \cdot x^{\hat{\beta}} \leftrightarrow log(y) = \hat{\alpha} + \hat{\beta} \cdot log(x)$$

When we increase $log(x)$ with 1 unit, we get:

$$log(y) = \hat{\alpha} + \hat{\beta}_i \cdot (log(x) + 1)$$

$$y = e^{\hat{\alpha}} \cdot e^{\hat{\beta}(log(x)+1)}$$

$$y = e^{\hat{\alpha}} \cdot (e^{log(x)} \cdot e^1)^{\hat{\beta}}$$

$$y = e^{\hat{\alpha}} \cdot x^{\hat{\beta}} \cdot e^{\hat{\beta}},$$

which means that the insurance claim increases with a factor $e^{\hat{\beta}}$ when log(x) increases with 1 unit. An increase of log(x) with one unit is equal to an increase of x with a factor $e(x \cdot e)$.

With the confidence interval for β we get the insecurity of $\hat{\beta}$. It is of great importance having a narrow confidence interval for β due to $\hat{\beta}$:s affect on the insurance claims.

As we can see from the table above, the $\hat{\beta}$-value for the Agricultural sector is higher in comparison with the other sectors. This would suggest that the Agricultural sector would be more sensitive to a change in temperature index. The $\hat{\beta}$-value for the Agricultural sector in Norrland is 2.41. This means that if the log-transformed temperature index would increase with 1 unit (which corresponds with the difference between the winter with the highest and the winter with the lowest temperature index) it is equal to an increase in temperature index with a factor $e(\approx 2.72)$ the insurance claims would increase with a factor $e^{2.41} \approx 11.13$. Götaland on the other hand has a $\hat{\beta}$-value of 1.65, why an increase in the log-transformed temperature index would result in an increase in insurance claims with 5,21.

In Götaland and in the Private sector in general, we seem to have a smaller insecurity due to a narrower 95% confidence interval for $\hat{\beta}$ in comparison to the other regions and sectors and we also get a higher R^2, why the model appears stronger for the Private sector.

6.3.3 Prediction Interval

The 95% prediction interval is an estimate of the interval within which new predicted insurance claims (y_0) would fall for every x_0. In this material, y_0 and x_0 have been log-transformed. The prediction interval line in the point $log(x_0)$ is given below:

$$\hat{\alpha} + \hat{\beta} \cdot log(x_0) \pm t_{p/2}(n-2)\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(log(x_0) - \overline{log(x)})^2}{\sum_{i=1}^{n}(log(x_i) - \overline{log(x)})^2}}$$

To simplify:

$$t_{p/2}(n-2)\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(log(x_0) - \overline{log(x)})^2}{\sum_{i=1}^{n}(log(x_i) - \overline{log(x)})^2}} = \delta$$

which gives that the 95% prediction interval:

$$\hat{\alpha} + \hat{\beta} \cdot log(x_0) \pm \delta$$

Since we have used a log-transformation, from the formula above we get that the 95% prediction interval's affect on y_0 is derived after multiplying or dividing the prediction value with e^{δ} as shown below:

$$0.95 = P(\hat{\alpha} + \hat{\beta} \cdot log(x_0) - \delta < log(y_0) < \hat{\alpha} + \hat{\beta} \cdot log(x_0) + \delta)$$

$$\iff$$

$$0.95 = P\left(\frac{e^{\hat{\alpha}} \cdot x_0^{\hat{\beta}}}{e^{\delta}} < y_0 < e^{\hat{\alpha}} \cdot x_0^{\hat{\beta}} \cdot e^{\delta}\right)$$

A wide prediction interval leads to insecurity for new observations in the model. Below we demonstrate the δ-values for an average winter season (2000-01) for each region and sector in order to demonstrate the size of δ, since we are not able to show any insurance claim values due to the confidentiality agreement:

- Norrland; Private: 0.54, Agricultural: 0.97, Commercial: 0.85
- Svealand; Private: 1.00, Agricultural: 1.44, Commercial: 1.53
- Götaland; Private: 1.00, Agricultural: 1.55, Commercial: 1.30

These values display a great insecurity due to a large spread. The δ-value for the Agricultural sector in Norrland is 0.97. To obtain the prediction interval for the "un-log-transformed" $\hat{y}$ we multiply and divide with a factor e^{δ}=2.66. For an estimated insurance claim of 1 for a given temperature index, this would result in a spread between 0.38 and 2.66, written (0.38,2.66).

The value of δ seems smaller in Norrland and within the Private sector. The largest spread is seen in the Agricultural sector, which was also true for the spread for the $\hat{\beta}$-values before.

6.4 Comparison between Minimum and Mean Temperature Index

When creating the temperature index, all temperatures above 0°C were removed. When using the daily minimum temperatures for our temperature index, days with a mean temperature above 0°C, but a minimum temperature below 0°C, will be included in the index. Also, the minimum temperatures are lower than the mean temperatures, giving us a higher temperature index. How will this affect the relationship between the temperature index and the insurance claims? Which one of minimum and mean is better to use to predict insurance claims? Here the ECA&D temperatures will be used, since we had access to their minimum temperature data. We will also use the ECA&D mean temperatures, since we have to use the same stations for both mean and minimum, when we are comparing the two. Scatter plots and simple linear regression analysis are made for both temperature indexes, first without log-transformation and thereafter with log-transformation. As shown before, the log-transformed variables gave better results. The log-transformed plots are shown in *Figures 9-10*.

Below we demonstrate scatter plots, for the Private sector for minimum and mean temperature index with log-transformation.

y_i = insurance claims
x_1 = mean temperature index
x_2 = minimum temperature index

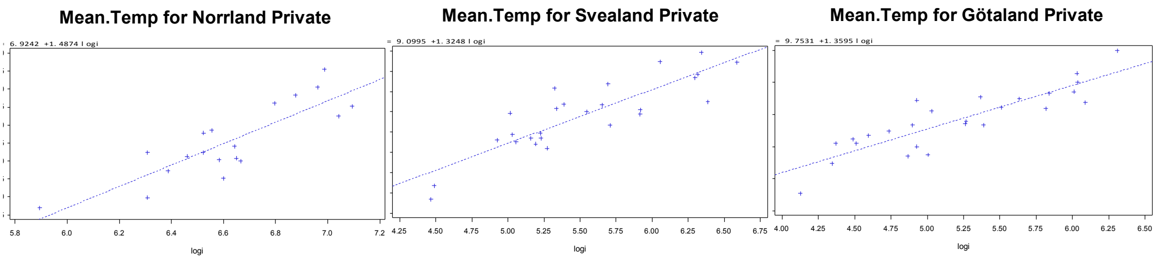


Figure 9: Scatter plots for the Private sector with log(y_i) and log(x_1) for Norrland, Svealand and Götaland.

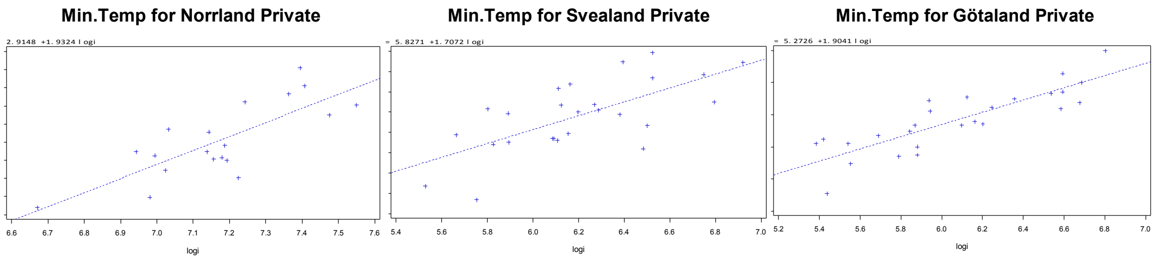


Figure 10: Scatter plots for the Private sector with log(y_i) and log(x_2) for Norrland, Svealand and Götaland.

In both the mean and the minimum temperature analysis, we see that the log-transformed variables decrease heteroscedacity as we saw previously. As explained before, since we use the ECA&D-dataset, fewer stations are used for the temperature indexes for this analysis. Even though fewer stations are used to calculate the index, there is still a strong correlation between mean temperature index and insurance claims.

The minimum temperatures are generally not as strongly correlated with the insurance claims as the mean temperatures. In *Table 5* we see correlation coefficient for the mean- and minimum temperature indexes. A further discussion regarding reasons for this will be found in the discussion section.

Correlation coefficient for private sector	*Mean Temperature index*	*Minimum Temperature index*
Norrland	0.81	0.78
Svealand	0.86	0.68
Götaland	0.87	0.85

Table 5: comparing the Private sectors correlation coefficient for mean- and minimum temperature index for Norrland, Svealand and Götaland.

6.4.1 Regression Analysis for Minimum and Mean Temperature Index

We have now seen that the mean temperature index has a stronger correlation. We continue with a simple linear regression in order to investigate further if the relationship with insurance claims differs between the two indexes.

The residuals for both the mean and minimum temperature index have been analyzed in the same way as in the section "Examine the Relationship". The residuals appear to be independent and random. With the Shapiro-Wilk W-test, we saw high p-values indicating that the data is normally distributed. We find no evidence of absence of normal distribution. The residuals have also been checked for autocorrelation with the Durbin-Watson test. For the minimum temperature indexes, we find that d for all regions and sectors ranges between 1,3 and 2,5 with a high p-value, indicating a low significance in the correlation between years and the residuals. The residuals for the mean temperature indexes were identical and varied between 1,3 and 2,5, with a high p-value, indicating non-autocorrelation.

The next step is to make a simple linear regression analysis, comparing the $\hat{\beta}$-values, 95% confidence interval, the coefficient of determination and significance. This is demonstrated in *Tables 6-11*.

Norrland Mean Temp.	$\hat{\beta}$-value	*95% Conf. for* $\hat{\beta}$	R^2	*p-value*
Private	1.49	(1.01, 1.96)	0.72	<0.0001
Agricultural	2.33	(1.58, 3.08)	0.72	<0.0001
Commercial	1.42	(0.70, 2.14)	0.50	0.0070

Table 6: Linear regression analysis with $log(y_i)$ and $log(x_1)$ from Norrland the with R^2-values, p-values and $\hat{\beta}$-values and their respective confidence from ECA&D

Norrland Min. Temp.	$\hat{\beta}$-value	*95% Conf. for $\hat{\beta}$*	R^2	*p-value*
Private	1.93	(1.15, 2.71)	0.62	<0.0001
Agricultural	3.15	(2.01, 4.29)	0.67	<0.0001
Commercial	1.85	(0.77, 2.93)	0.43	0.0022

Table 7: Linear regression analysis with $log(y_i)$ and $log(x_2)$ from Norrland the with R^2-values, p-values and $\hat{\beta}$-values and their respective confidence from ECA&D

Svealand Mean Temp.	$\hat{\beta}$-value	*95% Conf. for $\hat{\beta}$*	R^2	*p-value*
Private	1.32	(0.99, 1.65)	0.74	<0.0001
Agricultural	1.60	(1.12, 2.08)	0.66	<0.0001
Commercial	1.26	(0.75, 1.78)	0.52	<0.0001

Table 8: Linear regression analysis with $log(y_i)$ and $log(x_1)$ from Svealand the with R^2-values, p-values and $\hat{\beta}$-values and their respective confidence from ECA&D

Svealand Min. Temp.	$\hat{\beta}$-value	*95% Conf. for $\hat{\beta}$*	R^2	*p-value*
Private	1.71	(0.93, 2.48)	0.47	0.0001
Agricultural	2.14	(1.14, 3.14)	0.45	0.0002
Commercial	1.99	(0.94, 3.06)	0.39	0.0007

Table 9: Linear regression analysis with $log(y_i)$ and $log(x_2)$ from Svealand the with R^2-values, p-values and $\hat{\beta}$-values and their respective confidence from ECA&D

Götaland Mean Temp.	$\hat{\beta}$-value	*95% Conf. for $\hat{\beta}$*	R^2	*p-value*
Private	1.36	(1.04, 1.68)	0.76	<0.0001
Agricultural	1.73	(1.23, 2.22)	0.68	<0.0001
Commercial	1.12	(0.71, 1.53)	0.57	<0.0001

Table 10: Linear regression analysis with $log(y_i)$ and $log(x_1)$ from Götaland the with R^2-values, p-values and $\hat{\beta}$-values and their respective confidence from ECA&D

Götaland Min. Temp.	$\hat{\beta}$-value	*95% Conf. for $\hat{\beta}$*	R^2	*p-value*
Private	1.90	(1.41, 2.40)	0.72	<0.0001
Agricultural	2.36	(1.57, 3.14)	0.61	<0.0001
Commercial	1.59	(0.99, 2.19)	0.56	<0.0001

Table 11: Linear regression analysis with $log(y_i)$ and $log(x_2)$ from Götaland the with R^2-values, p-values and $\hat{\beta}$-values and their respective confidence from ECA&D

In all regions there is a significant relationship between insurance claims and both mean and minimum temperature index at a 5% level. This is seen through a p-value less than 0.0001.

The $\hat{\beta}$-value is smaller for the mean temperature index and the 95% confidence intervals are narrower than for the minimum temperature indexes. We can also see that the coefficient of determination is higher for the mean temperature index. Hence, the mean temperature index appears to have a stronger relationship with insurance loss than does the minimum temperature index. Through the $\hat{\beta}$-value, we also see a bigger effect on the insurance claims with an increase in the mean temperature index.

We also study the prediction interval. A wide prediction interval leads to insecurity in the model. Below we demonstrate the "δ-values" for an average winter season (1996-97)– as shown in section "Examining the Relationship" – for each region and sector in order to demonstrate the size of δ:

- Norrland_{Mean}; Private: 0.61, Agricultural: 0.96, Commercial: 0.93
- $\text{Norrland}_{Min.}$; Private: 0.72, Agricultural: 1.05, Commercial: 0.99
- Svealand_{Mean}; Private: 0.96, Agricultural: 1.41, Commercial: 1.51
- $\text{Svealand}_{Min.}$; Private: 1.41, Agricultural: 1.83, Commercial: 1.93
- Götaland_{Mean}; Private: 1.03, Agricultural: 1.59, Commercial: 1.31
- $\text{Götaland}_{Min.}$; Private: 1.11, Agricultural: 1.76, Commercial: 1.33

Mean temperature index has lower values for δ in all regions and sectors, which indicates more security in the model.

We conclude that the mean temperature index is a better explanatory variable for the insurance claims, since we have a smaller 95% confidence interval for $\hat{\beta}$, a more narrow prediction interval and a higher R^2.

6.5 Multiple Linear Regression

Multiple linear regression analysis was tested on the model. A log-transformed model has been used:

$$log(y_i) = \hat{\alpha} + \hat{\beta}_1 \cdot log(x_{1,i}) + \hat{\beta}_2 \cdot log(x_{2,i})$$

y_i = insurance claim
$\hat{\alpha}$ and $\hat{\beta}$ = estimated parameters
x_1 = mean temperature index
x_2 = minimum temperature index
ε_i = an independent and normally distributed N(0,σ^2) error term.

The coefficient of determination is increased in comparison with the simple linear regression model, although this was expected, since R^2 increases with increasing number of variables.

With multiple linear regression analysis we could not se any definite improvement, but in some regions and sectors we could see significant relationships, which can be studied further. As reported before, the mean temperatures have been weighted with consideration to the minimum temperatures, why the minimum and mean temperatures will be dependent. We could see that in most cases the minimum temperatures were not significant. This will not be studied further in this thesis.

6.6 Summary of the Comparison between Minimum and Mean Temperatures

In conclusion, when comparing the scatter plots, the $\hat{\beta}$-value and the widths of $\hat{\beta}$'s 95% confidence intervals, 95% prediction intervals and the coefficients of determination, we conclude that the mean temperature indexes show a stronger relationship to the insurance claims than do the minimum temperature indexes. Consequently from now on, we will only analyze the mean temperature indexes.

7 Prediction of Insurance Claims 1930-2011

As mentioned before, the winters the past two decades have been quite warm in comparison to the winters in the past century. In order to look at the relationship for the colder winters that have occurred the past 80 years we have to conduct a prediction analysis of the material. SMHI has daily temperatures between 1961 and 2011. In order to make predictions regarding colder winters, we want to include the period 1939-1942, why the SMHI-data is not sufficient. In the ECA&D-dataset, some stations have data going back all the way to 1875, although we are only interested in analyzing the period 1930-2011. *Figure 11-13* demonstrate the temperature indexes between 1930-2011 for Norrland, Svealand and Götaland.

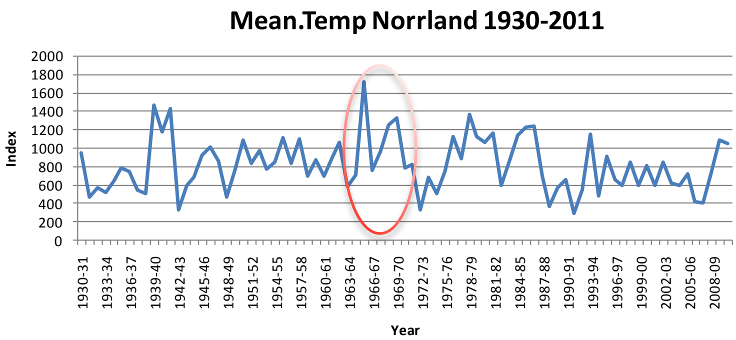


Figure 11: Mean temperature index for Norrland between 1930-2011, ECA&D

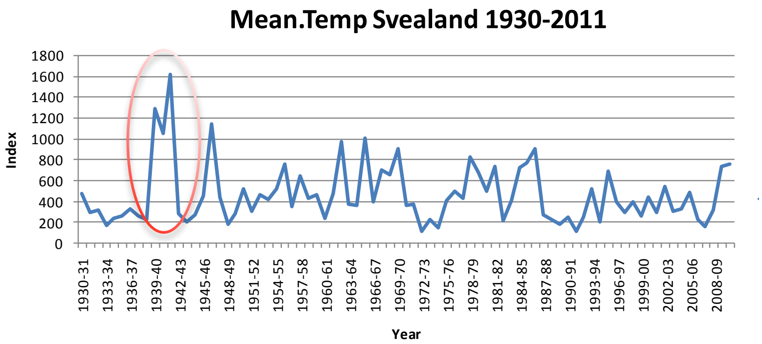


Figure 12: Mean temperature index for Svealand between 1930-2011, ECA&D

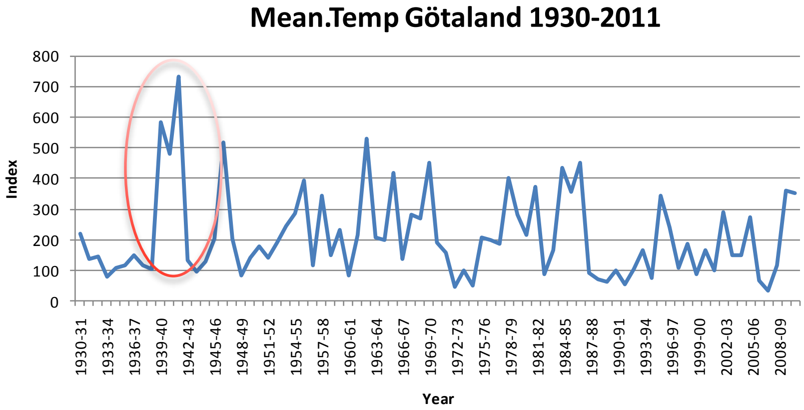


Figure 13: Mean temperature index for Götaland between 1930-2011, ECA&D

There were only four ECA&D temperature stations that had data from 1930-2011, one station in Norrland, two in Götaland and Köpenhamn. We included Köpenhamn as a part of Götaland to represent the southern part of Götaland. We found one station in Svealand with data going back to 1950.

The data from the three stations in Götaland was aggregated and then weighted with regard do the insurance claims in the regions as preformed the section “Weighting the temperature index”.

The stations used in each region are presented in the table below.

Region	*Temperature Station*
Norrland	Frösön
Svealand	Karlstad Airport
Götaland	Köpenhamn, Växjö and Linköping

Table 12: Temperature station used per region

We constructed scatter plots where we compared the temperature index from each region with the other two respectively. They were all checked for correlation, and we noticed that the temperatures were strongly correlated through a linear relationship.

Through a correlation matrix between Götaland and Svealand, we could se a strong positive correlation ($r = 0{,}96$). To point estimate temperatures for Svealand we took the quota from dividing the Svealand temperature index with the Götaland temperature index. We found that Svealand had an average temperature index of 2,21 times the average Götaland temperature index. Therefore, through this correlation, we could point estimate the 20 years from 1930-1950 for the station in Svealand.

Now that we have new mean temperature indexes for all three regions for 1930-2011, we once again must conduct a simple linear regression analysis with log-transformation. Since we only have insurance claim data between 1985 and 2011, the linear regression will be made with the temperatures from those years. *Figure 14* demonstrates log-transformed values for the Private sector in all regions.

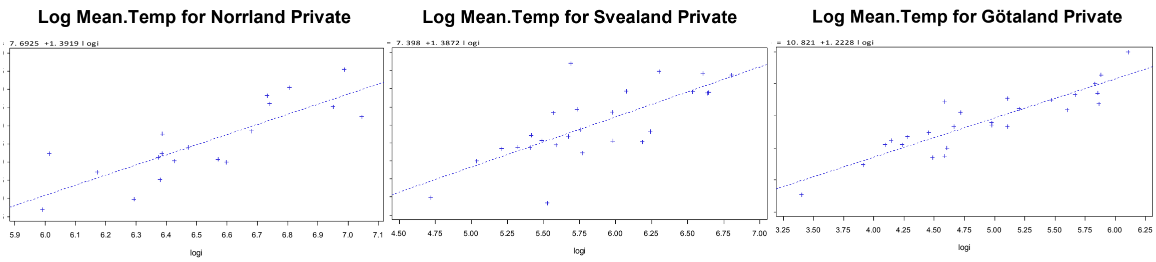


Figure 14: Scatter plots from Norrland, Svealand and Götaland – Private sector: log(yi) in relation to log(xi) with regressions line. Correlation coefficient – Norrland: r =0.82, Svealand: r =0.86 , Götaland: r = 0.90

The residuals have been analyzed in the same way as in the section "Examine the Relationship" and the section "Comparison between minimum and mean". The residuals appear to be independent and random. The Shapiro-W-test gives high *p*-values indicating that the data is normally distributed. We find no evidence of absence of normal distribution. The residuals have also been checked for autocorrelation with the Durbin-Watson test. We find that *d* for all regions and sectors ranges between 1,5 and 2,5 with high *p*-values indicate a low significance for the autocorrelation.

The results from the regression analysis is presented in *Table 13-15*.

Norrland Mean Temp.	$\hat{\beta}$-value	*95% Conf. for* $\hat{\beta}$	R^2	*p-value*
Private	1.39	(0.89, 1.89)	0.67	<0.0001
Agricultural	2.16	(1.37, 2.96)	0.66	<0.0001
Commercial	1.46	(0.81, 2.11)	0.57	0.0002

Table 13: regression summary with log(y_i) and log(x_i) from mean temperature station Frösön from Norrland

Svealand Mean Temp.	$\hat{\beta}$-value	*95% Conf. for* $\hat{\beta}$	R^2	*p-value*
Private	1.41	(1.06, 1.77)	0.74	<0.0001
Agricultural	1.68	(1.15, 2.11)	0.64	<0.0001
Commercial	1.39	(0.85, 1.92)	0.57	<0.0001

Table 14: regression summary with log(y_i) and log(x_i) from mean temperature station Karlstad from Svealand

Götaland Mean Temp.	$\hat{\beta}$-value	*95% Conf. for* $\hat{\beta}$	R^2	*p-value*
Private	1.23	(0.97, 1.48)	0.80	<0.0001
Agricultural	1.58	(1.20, 1.97)	0.75	<0.0001
Commercial	0.99	(0.63, 1.34)	0.58	<0.0001

Table 15: regression summary with log(y_i) and log(x_i) from mean temperature station Köpenhamn, Växjö and Linköping from Götaland

In all regions and sectors, there is a significant relationship between insurance claims and the temperature index at a 5% level. This is seen through a p-value less than 0.0001. R^2 is higher for the Private sector than for the Agricultural- and Commercial sector. As we saw before in the section "Examining the Relationship" the $\hat{\beta}$-values are higher in the Agricultural sector. This indicates that an increase in temperature index has a bigger effect on the insurance claims from the Agricultural sector than from the other sectors. The $\hat{\beta}$-interval for the private sector has the narrowest 95%-confidence interval for $\hat{\beta}$, which decreases the insecurity. The low $\hat{\beta}$-value indicates that an increase in temperature index has a smaller effect on the insurance claims from the Private sector.

Agriculture still has the greatest uncertainty among the sectors. Generally, Svealand has the greatest uncertainty among the regions.

Through the regression analysis, we have found the models to be significant at a 5% level with the constant of determination at a varying level. We cannot prove that ε_i is not normally distributed. With this in mind, we now proceed to "predict" the period 1930-1985.

In order to estimate new values for the insurance claims for the period 1939-1942, we use the prediction method and use the 95% prediction interval to get an estimation of the variation around the regression line. The prediction of the insurance claims for Götaland is shown below with insurance loss excluded due to the signed confidentiality agreement.

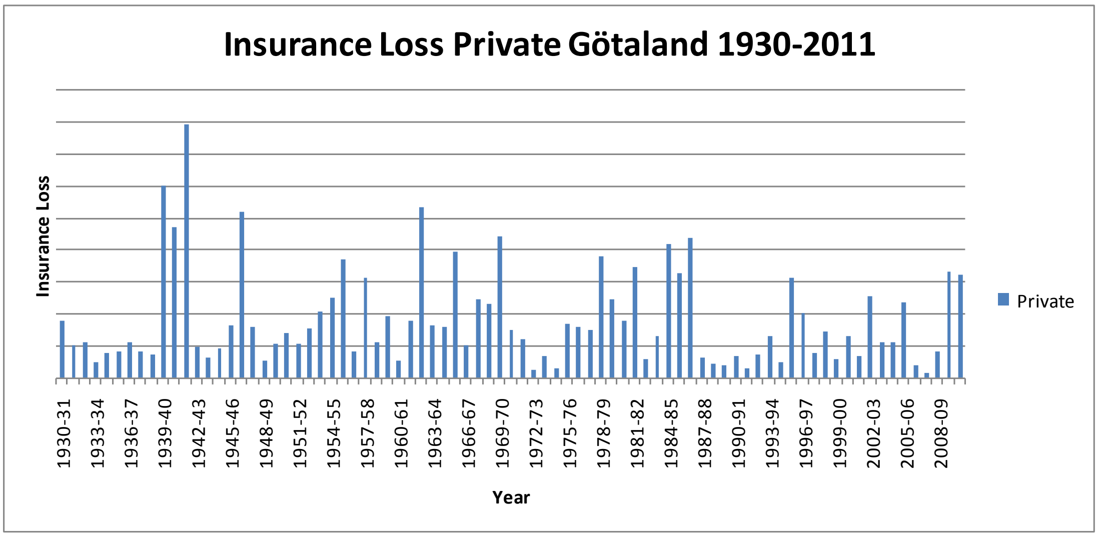


Figure 15: Predicted Insurance loss for the Private sector in Götaland between 1930-1985

Figures for the rest of the region and sectors will be found in Appendix II. From this figure, we get that if a winter as cold as the winters 1939-1942 would occur, the insurance claims would be increase considerably, which will be discussed further on the next page. As diagnostic methods of model criticism we use the prediction intervals to examine the spread around $\hat{y}_i$.

In *Figure 16*, we see the log-transformed insurance claims for the Private sector from the original dataset between 1985-2011 on the y-axis. On the x-axis are the log-transformed predicted insurance claims from the same period. The reason for constructing such a plot is to investigate the relationship and to evaluate the predicted values. It appears as if the if there is a linear relationship between the two, which supports the prediction.

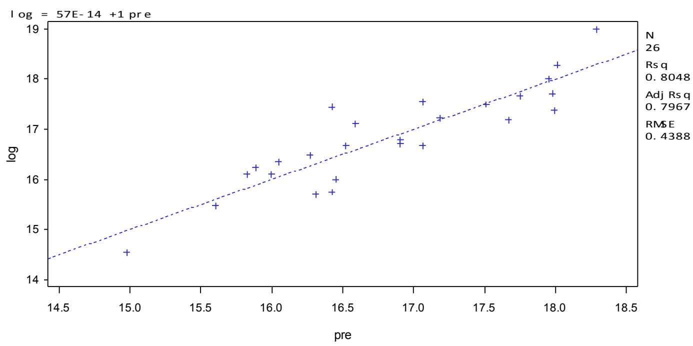


Figure 16: Private insurance claims from Götaland for year 1985-2011: log-transformed values from the

original dataset over log-transformed the log-transformed predicted dataset.

To illustrate the meaning of the prediction interval, we will demonstrate the effect of the prediction interval for an "average" winter, the coldest winter within the period with insurance data and the coldest winter within the prediction as an "out of sample prediction" (*Table 16*). The "e^{δ}-values" for the prediction intervals have been calculated as shown in section "Examining the Relationship". Below the temperatures from Götaland and the "e^{δ}-values" are presented. e^{δ} represents the quota between y_0 and the lower limit if the prediction interval. It is also the quota between the upper limit of the prediction interval and y_0. Tables for Norrland and Svealand will be presented in Appendix III.

Götaland/ Winter	*2001-02 "average"*	*1986-87 coldest with insurance data*	*1941-42 coldest within prediction*
Temperature Index	98	449	731

Table 16: Temperature index for Götaland during an "average" winter, the coldest winter within the period with insurance data and the coldest winter within the prediction

If we assume that the insurance claims for the Private sector in Götaland for the "average" winter 2001-2002 would be 1 unit, the corresponding claims for the other two winters are demonstrated. As we have shown in the section "Examining the Relationship", we will now show the e^{δ}-values' effect on the prediction intervals for the periods above. To obtain the prediction interval for the "un-log-transformed" $\hat{y}$, we multiply/divide with a factor e^{δ} for each period. The e^{δ}-values are demonstrated in *Table 17* and the prediction intervals effect on the insurance claims in Götaland is demonstrated in *Table 18*.

Götaland	*Winter 2001-02*	*Winter 1986-87*	*Winter 1941-42*
Private	2.53	2.59	2.77
Agricultural	4.06	4.22	4.66
Commercial	3.36	3.78	4.14

Table 17: Götaland- " e^{δ}-values" for an "average" winter, the coldest winter within the period with insurance data and the coldest winter within the prediction

Götaland	*Winter 2001-02*	*Winter 1986-87*	*Winter 1941-42*
Private	1	6.43	11.67
Pred. intervall for Private	(0.39, 2.53)	(2.48, 16.65)	(4.21, 32.33)
Agricultural	1	11.16	24.67
Pred. intervall for Agricultural	(0.25, 4.06)	(2.64, 47.10)	(5.29, 114.96)
Commercial	1	4.49	7.26
Pred. intervall for Commercial	(0.28, 3.63)	(1.19, 16.97)	(1.75, 30.06)

Table 18: Predicted insurance claims from the Private, Agricultural and Commercial sector in Götaland if the claims for the Winter 2001-02 would be 1 unit. Second row shows the prediction intervals' effect on $\hat(y)$ if $\hat(y)$ for the winter 2001-02 would be 1 unit.

We see in *Table 17* that the predictions for the Private sector have narrower prediction intervals than do the other sectors. In *Table 18* we show that the prediction from the model is very uncertain, which is demonstrated with a wide prediction interval. The "out of sample prediction" has much wider prediction intervals than do the other predictions. Although, we can say that a winter, as cold as the winter 1941-42, would generate big insurance claims in comparison with an average winter from the past two decades.

7.1 Predication for Sweden as a Whole

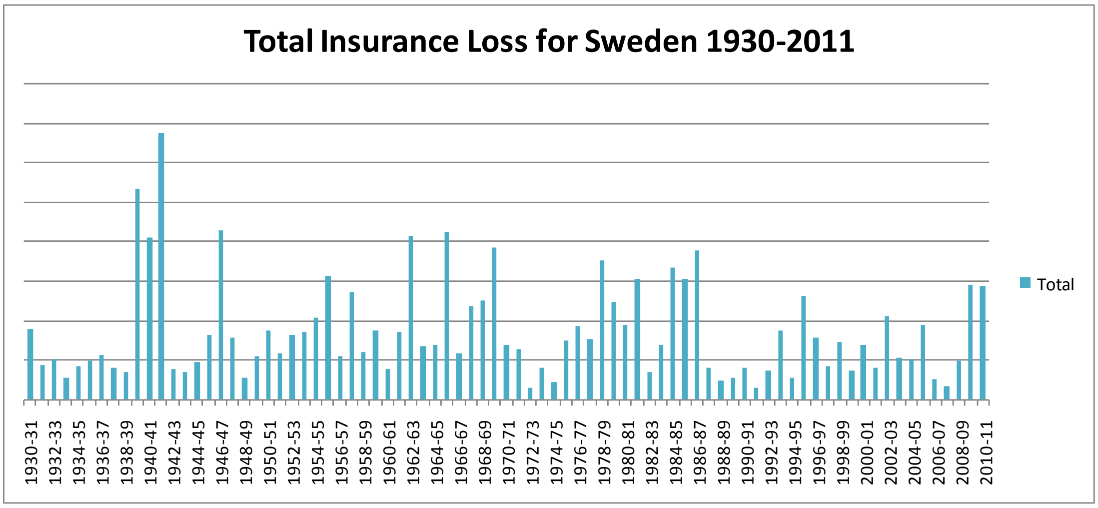


Figure 17:Predicted Total Insurance loss for Sweden between 1930-1985

We construct a prediction for the period 1930-2011 with the data from all sectors and regions aggregated into a whole, i.e. the insurance claims from Sweden as a whole. The prediction is made under the assumption that all sectors and regions are independent of each other and thereby summed the predicted insurance claims.

What can be seen from the prediction is that the past two winters should have been among the 15 winters with the highest insurance claims of the past 80 years. We can state with great uncertainty that a winter, as cold as the winter 1941-42, would probably have generated higher insurance claims than the winters of 2009-2011.

8 Results and Conclusion

Initially, the relationship between the temperature index and the claims from the each area was analyzed with scatter plots, but no relationship was seen. We also noted that the claims from these areas were not large enough to show any relationship with the temperature data, why the insurance data had to be aggregated and the areas merged into larger regions.

The areas were grouped into three major regions of Sweden: Norrland, Svealand and Götaland. 18.59% of the insurance company's total claims of the period were from Norrland, 22.13% from Svealand and 59.28% from Götaland.

In Sweden as a total, the distribution of insurance claims between the three sectors was as follows: Private 46%, Agriculture 12% and Commercial 42%.

In all analyses for the mean temperature index from the SMHI dataset, the mean and the minimum temperature indexes from the ECA&D dataset, we could see that the log-transformed variables decreased the heteroscedacity. The minimum temperatures were generally not as strongly correlated with the insurance claims as the mean temperatures, why the mean temperatures were used for the analysis.

Through a simple linear regression analysis with the insurance data over the temperature index derived from the SMHI temperature data we could see that the $\hat{\beta}$-value for the Agricultural sector was higher in comparison with the other sectors. We could also see that the $\hat{\beta}$-value for Norrland was higher in comparison with the other regions. This would suggest that the Agricultural sector and Norrland as a region would be more sensitive to a change in temperature index than would the other sectors and regions. As an example, the $\hat{\beta}$-value for the Agricultural sector in Norrland was 2.41. This would mean that if the log-transformed temperature index were to increase with 1 unit (the difference between the year with the highest and the lowest temperature index) the insurance claims would increase with a factor $e^{2.41} \approx 11.13$.

However, in Götaland and in the Private sector, the insecurities were smaller owed to a narrower 95% confidence interval for $\hat{\beta}$ in comparison to the other regions and sectors. We also got a higher R^2, why the model appeared stronger for the Private sector.

Through the method of prediction we could construct graphs with insurance claims for the period 1930-2011 with wide prediction intervals suggesting uncertainty in the model. Although, we can say that a winter, as cold as the winter 1939-42, would generate big insurance claims in comparison with an average winter from the past two decades.

9 Discussion

The purpose of this thesis has been to examine how the freezing temperatures during wintertime affect the insurance claims on property during the same period.

To achieve this goal, we had to construct a model of numerically describing the coldness of a winter period. This was achieved with the help of Guy Carpenter although several other options of arbitrary numerical models could have been considered and tested. As we wrote in the section "Aims and Methods" we have taken two factors into consideration that probably is involved in damages caused by freezing.

1. The temperature (the number of degrees Celsius below freezing point – 0°C)
2. The "time-factor" i.e. number of days with the temperature above.

These two factors can be accounted for in many different arbitrary models of constructing a temperature index. As an example, we could look at the number of days below 0°C separately. It is difficult to comment on beforehand what model will have the strongest relationship with the insurance claims.

The minimum temperatures were generally not as strongly correlated with the insurance claims as the mean temperatures, why the mean temperatures were used for the analysis. One could argue that the reason for this would be that the temperature stations take the minimum temperature into consideration when calculating the mean temperature. The daily mean temperatures were estimated with several measurements over the day and thereby probably had more of the "time-factor" mentioned above accounted for, why the time below freezing might have a stronger influence than we had considered from the start.

Through simple linear regression, we have been able to construct a statistical model describing a strong correlation between the mean temperature index and insurance claims.

With multiple linear regression analysis we could not se any definite improvement, but in some regions and sectors we could see significant relationships, which can be studied further. One could also construct a model taking other temperature indexes and/or more environmental factors into consideration, such as snow depth and snow pressure and to look at possible interactions.

The reason to look at snow depth and pressure would be that snow pressure can be a direct cause of damages, but can at the same time insulate and hence minimize the effect of freezing.

When looking at the different sectors separately, we saw that the $\hat{\beta}$-value for the Agricultural sector was higher in comparison with the other sectors. This would suggest that the Agricultural sector would be more sensitive to a change in temperature index than the other sectors, which would be more intuitive due to a more natural "exposure" to the environment in the Agricultural sector. Though one must be careful drawing such conclusions, since the 95% confidence intervals are wide and overlap those of the other sectors. Small data samples in Norrland and for the Agricultural sector might also be a source of uncertainty.

We have chosen to look at the different regions and sectors (Private, Agricultural and Commercial) separately to see if there is a difference.

One aim with the thesis was to examine how big the claims would be today if the winters would be as cold as the coldest period this century. We have shown that a winter, as cold as the winter 1939-42, would generate big insurance claims in comparison with an average winter from the past two decades. It appears, using a 95% prediction interval (as shown in *Table 18*), that a winter as cold as the winter 1941-1942 would generate insurance claims from the Private sector somewhere between 4,2 and 32 times the insurance claims from an average winter.

One problem constructing the temperature index for the prediction was that we were forced to use the ECA&D-dataset, since they provided temperatures from before 1961 as opposed to SMHI. When we compare the R^2 for the correlation between the insurance claims and the temperature indexes constructed with the SMHI- and the ECA&D-data respectively, we see a higher R^2 for the SMHI-data, suggesting that the results would have improved if we had access to more temperature stations with data going back before 1961. Another problem using the ECA&D dataset was that we had to point estimate the temperatures for Svealand for the period 1930-1950, which increased the uncertainty of the prediction in the region of Svealand. We should even have considered omitting Svealand in the prediction analysis due to this fact.

Another source of insecurity in the method is that we have made an "out of sample prediction". The highest temperature index for which we have insurance data in Götaland was an index of 449, while we are trying to predict the insurance claims when the temperature index was 731. We noted that the difference in temperature index between our coldest "out of sample prediction" and the coldest index within the period with insurance data was largest in Svealand, corresponding with the widest prediction interval for the prediction in that region.

From the formula for the prediction interval:

$$log(\hat{y}_i) = \hat{\alpha} + \hat{\beta}(x_0) \pm t_{p/2}(n-2)\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}$$

we get that one way of decreasing the width of the prediction interval would be to increase the number of observations (n), in other words to retrieve more insurance data. The simplest way of doing this would be to merge the different sectors or to get insurance data from other insurance companies.

In order to confirm the results above, it would be preferable to construct a cross-validation study in order to see how well the results would work applied to the new dataset.

10 References

Books

Alm S. E. Och Britton T. (2008). Stokastik – sannolikhetsteori och statistiskteori med tillämpningar. Stockholm. Liber AB

Andersson P. & Tyrcha J. (2010)-Kompendium

Gustafsson B. (2000). Återförsäkring. Tierps Tryckeri AB/Partner Print

Løvås G. (2006). Statistik – metoder och tillämpningar. Liber

Ohlsson, E (2005). Kort handledning i SAS. Matematiska institution, Stockholm universitet

Sundberg, R. (2011). Lineära Statistiska Modeller. Matematiska institution, Stockholms universitet

Vejde O. and Leander E. (2000). Ordbok I statistic – ord I rättan tid såsom gyllene äpplen uti silverskålar. Ordspråksbok 25:11, Bibeln. Olle Vejde Förlag

Internet

http://www.smhi.se/omsmhi/Om-SMHI/vad-gor-smhi-1.8125 2012-04-21 kl. 13.11

http://eca.knmi.nl/ 2012-04-30 kl. 16.00

http://www.guycarp.com/portal/extranet/index.html?vid=84 2012-05-11 kl.16.43

http://support.sas.com/ 2012-05-14 kl. 19.07

http://www.jmp.com/support/faq/jmp2085.shtml 2012-05-18 kl. 20.43

11 Appendix

Appenix I

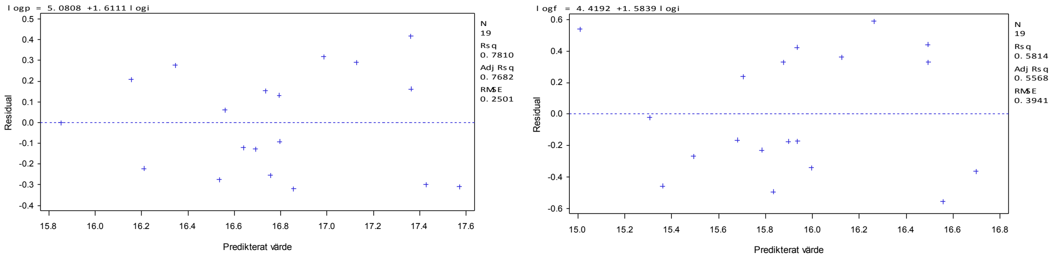


Figure 18: Norrland- Residual over prediction for Private and Commercial

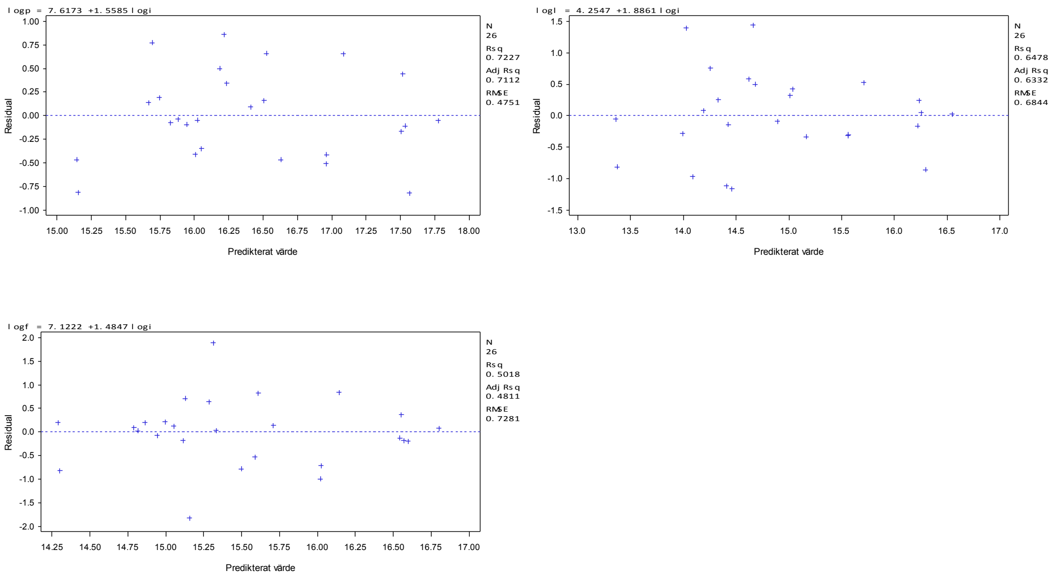


Figure 19: Svealand- Residual over prediction for Private, Agricultural and Commercial

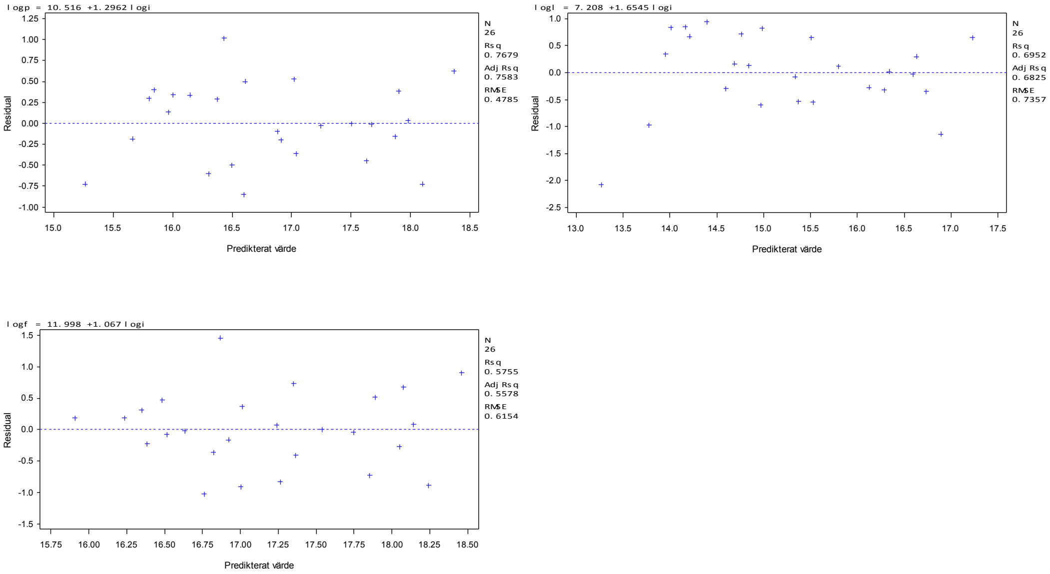


Figure 20: Götaland- Residual over prediction for private, Agricultural and Commercial

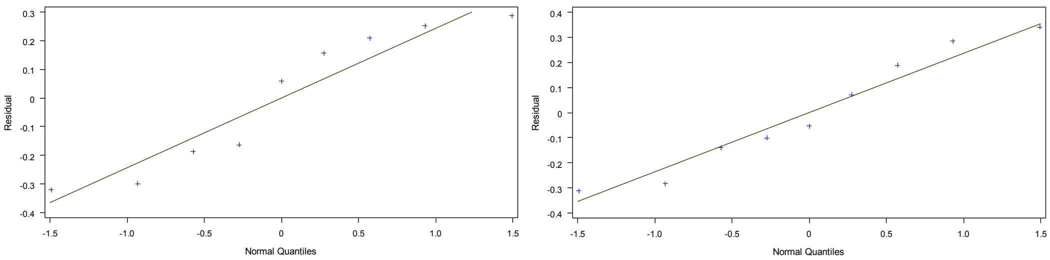


Figure 21: QQ-plots over residual Norrland: Private and Commercial

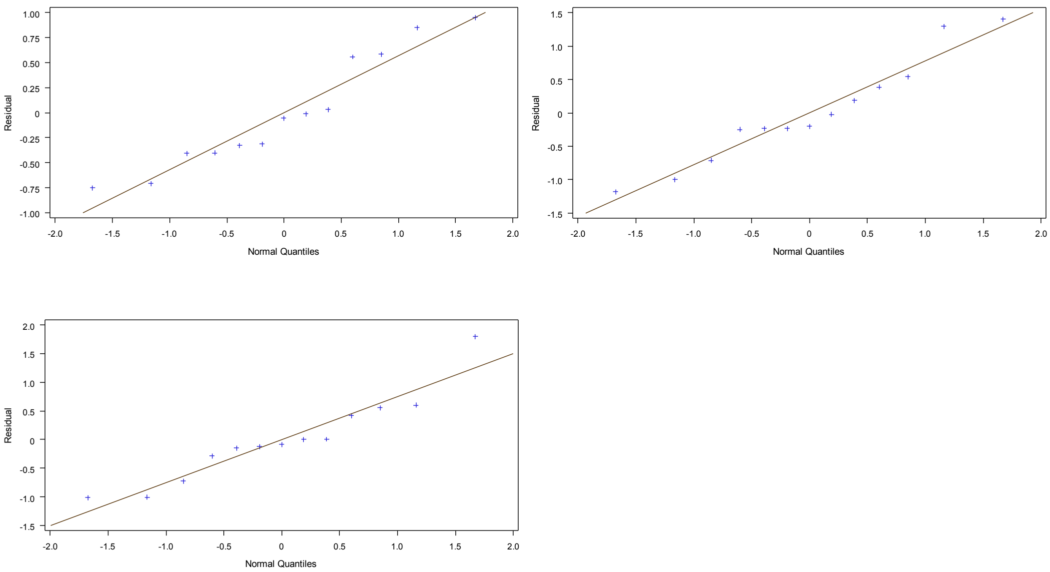


Figure 22: QQ-plots over residual Svealand: Private, Agricultural and Commercial

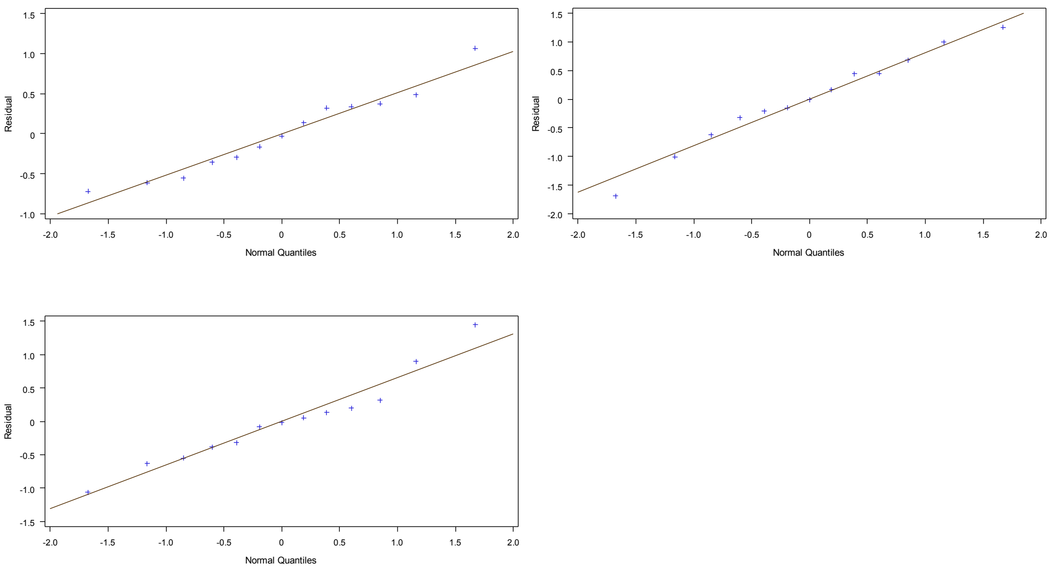


Figure 23: QQ-plots over residual Götaland: Private, Agricultural and Commercial

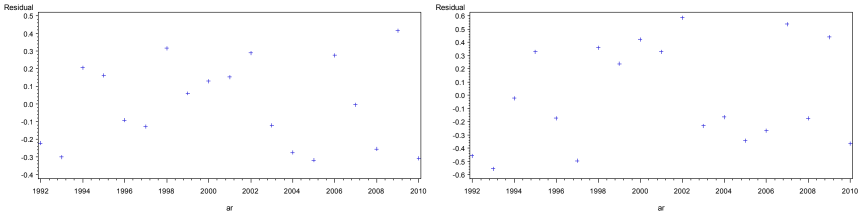


Figure 24: Norrland: Residual over year- Private and Commerciall

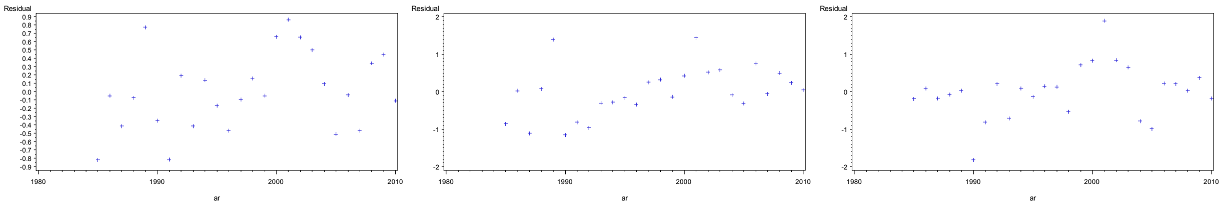


Figure 25: Svealand: Residual over year - Private, Agricultural and Commercial

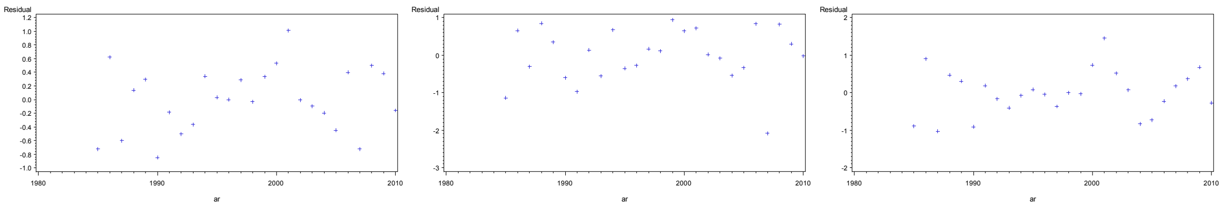


Figure 26: Götaland: Residual over year- Private, Agricultural and Commercial

Autocorrelation coefficient	*Private*	*Agricultural*	*Commercial*
Norrland	0.048	-0.14	0.09
Svealand	0.16	0.008	0.33
Götaland	-0.16	-0.25	0.002

Table 19: Durbin-Watson Statistic for autocorrelation: Autocorrelation coefficient (-1<autocorrelation<1) for Private, Agricultural and Commercial for SMHI.

P-value for Autocorrelation	*Private post./neg*	*Agricultural post./neg*	*Commercial post./neg*
Norrland	0.22/0.78	0.52/0.48	0.16/ 0.84
Svealand	0.10/0.90	0.36/0.64	0.11/0.89
Götaland	0.67/0.33	0.81/0.19	0.33/0.67

Table 20: Durbin-Watson Statistic for autocorrelation: Positive- and negative P-value for Autocorrelation for SMHI

Appendix II

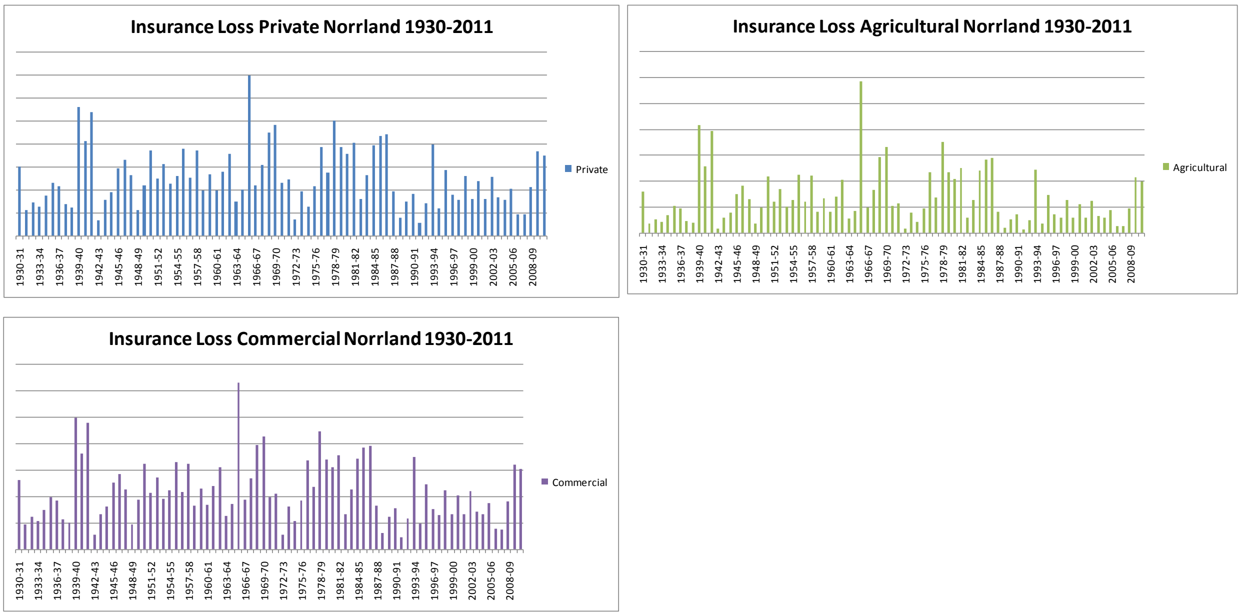


Figure 27: Insurance loss for the Private, Agricultural and Commercial sector in Norrland over time with point estimates of insurance loss between 1930-1985

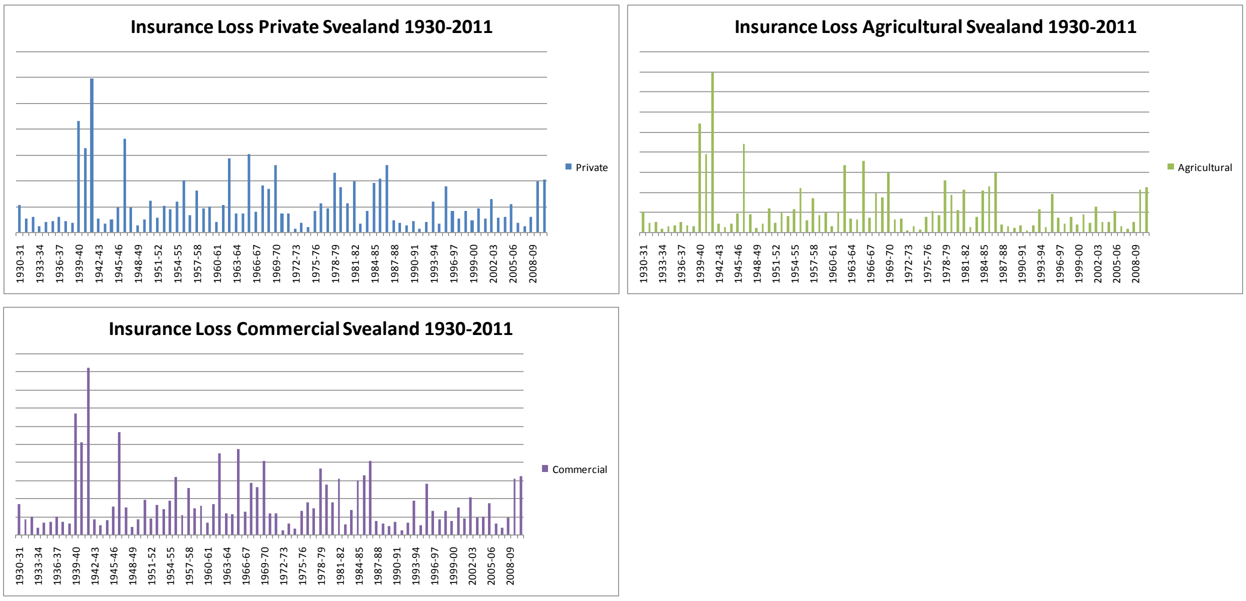
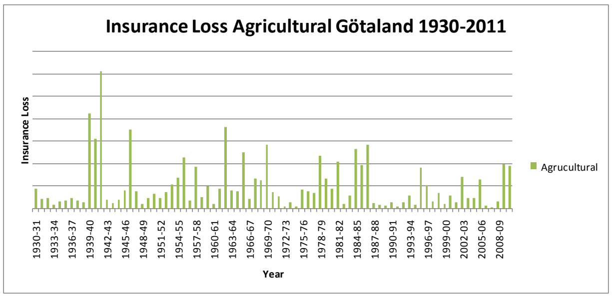


Figure 28: Insurance loss for the Private, Agricultural and Commercial sector in Svealand over time with point estimates of insurance loss between 1930-1985



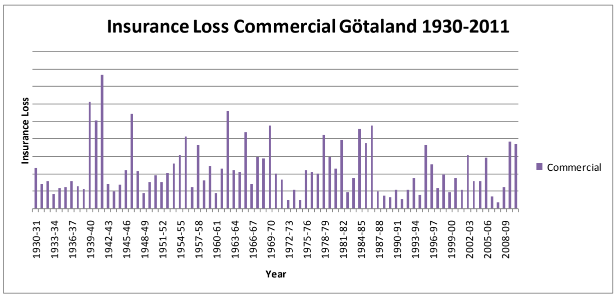


Figure 29: Insurance loss for the Agricultural and Commercial sector in Götaland over time with point estimates of insurance loss between 1930-1985

Appendix III

Norrland/ Winter	*2001-02 "average"*	*1993-94 coldest with insurance data*	*1965-66 coldest within prediction*
Temperature Index	594	1147	1720

Table 21: Temperature index for Norrland during an "average" winter, the coldest winter within the period with insurance data and the coldest winter within the prediction

Norrland	*Winter 2001-02*	*Winter 1993-94*	*Winter 1965-66*
Private	1.95	2.03	2.25
Agricultural	2.39	2.53	3.60
Commercial	2.89	3.10	2.86

Table 22: Norrland- " e^{δ}-values" for an "average" winter, the coldest winter within the period with insurance data and the coldest winter within the prediction

Svealand/ Winter	*2001-02 "average"*	*1986-87 coldest with insurance data*	*1939-40 coldest within prediction*
Temperature Index	296	902	1285

Table 23: Temperature index for Svealand during an "average" winter, the coldest winter within the period with insurance data and the coldest winter within the prediction

Svealand	*Winter 2001-02*	*Winter 1986-87*	*Winter 1939-40*
Private	2.64	2.77	2.94
Agricultural	4.31	4.66	5.00
Commercial	4.35	4.71	5.05

Table 24: Svealand- " e^{δ}-values" for an "average" winter, the coldest winter within the period with insurance data and the coldest winter within the prediction