Large claims in non-life insurance

Oscar Hagsjö & Oscar Hermansson

Acknowledgement

We would like to give big thanks to Sascha Firle at Trygg--Hansa for this amazing opportunity and great learning experience, and especially Emma Södergren for her support throughout the thesis and her valuable ideas and feedback.

We would also like to thank our mentor Boualem Djehiche for pointing us in the right direction when it was needed.

Abstract

It is of outmost importance for an insurance company to apply a fair pricing policy. If the price is too high, valuable customers are lost to other insurance companies while if it's too low – it nets a negative profit.

To achieve a good pricing policy, information regarding claim size history for a given type of customer is required. A problem arises as large extremal events occur and affects the claim size data. These extremal events take shape in individually large claim sizes that by themselves can alter the distribution for what certain groups of individuals are expected to cost.

A remedy for this is to apply what is called a large claim limit. Any claim exceeding this limit is thought of as being outside the scope of what is captured by the original distribution of the claim size. These exceeding claims are treated separately and have their cost distributed across all insurance takers, rather than just the group they belong to.

So, where exactly do you draw this limit? Do you treat the entire claim size this way (exclusion) or just the bit that is exceeding the threshold (truncation)?

These questions are treated and answered in this master's thesis for Trygg-Hansa.

For each product code, a limit was achieved in addition to which method for exceeding data that was best to use.

Storskador inom skadeförsäkring

Sammanfattning

Det är oerhört viktigt för ett försäkringsbolag att kunna tillämpa en god prissättning. Är priset för högt så förloras kunder till andra försäkringsbolag, och är den underprisad är det en förlustaffär.

För att kunna sätta bra priser krävs information om vilka samt hur stora skador som kan tänkas inträffa för en given kundprofil. Ett problem uppstår när stora extremfall påverkar skadedatan. Dessa extremfall yttrar sig genom enskilda storskador som kan komma att påverka prissättningen för en hel grupp då distributionen för vad gruppen förväntas kosta kan ändras.

Detta problem kan lösas genom att införa en storskadegräns till skadedatan. Skador över denna gräns räknas som extremfall och utanför ramen av vad den ursprungliga distributionen för skadorna beskriver. De hanteras separat och låter sin kostnad fördelas över samtliga försäkringstagare.

Men vart dras denna gräns? Ska man behandla hela den överstigande kostnaden på detta sätt (exkludering) eller bara den biten av skadan som går över storskadegränsen (trunkering)?

Dessa frågor behandlas och besvaras i denna masteruppsats i uppdrag åt Trygg-Hansa.

För de olika produkttypkoderna beräknades varsin storskadegräns samt metod för överskridande data.

Table of Contents

1 Introduction

The pricing of insurance for let's say a person A, is based on historical data of people with similar attributes. Given a man of a certain age and location, his insurance premium depends on what people just like him are expected to cost for the insurer. A problem arises when an extreme value or an outlying data point greatly affects this historical data.

For instance, if a person of the same measured attributes as person A has a very rare accident where the net cost is much larger than the average accident cost, this one incident will have significant effect on the premium for people who share the same attributes. Since this rare occasion probably could happen to anyone and should not necessarily be captured by the people of those attributes (which would result in their premiums being over-priced), the remedy is to remove this value from the data.

However, you cannot just remove this data point and forget about it since the fact that this significant but rare event still occur and is a very real cost for the insurance company. Therefore, you take that extreme cost and split it up between all different groups so that one set of people with the same attributes does not take the entire blow.

So, which data points are considered extreme enough to process this way?

- These points are referred to as Large claims and are defined by exceeding a certain threshold. The objective of this thesis is to find this limit and determine whether the large claims should be excluded or truncated as they are split across all insurance holders.

A large data set of modified insurance claims is supplied by Trygg-Hansa and is the foundation of this thesis. The problem is broken down and treated in two separate steps. The first step is to find the actual large claim limit. This is done with various mathematical methods that identify and analyze extreme value data. Once a limit has been obtained, the second step is to determine whether truncation or exclusion is best for treatment of the large claims.

The reader will first encounter a background which gives a brief introduction to non-life insurance as well as a necessary mathematical foundation of different expressions and methods used in the thesis. A description of the data set that was supplied can also be found in the background. Due to secrecy, the entirety of the data could not be included.

With a solid understanding of the methods and how they work, the puzzle of putting it all together to obtain the desired results are provided in the third chapter, *Execution & Result.* The result is then found in sub section of that chapter, *result*. Due to the nature of the methods there are a lot of graphical results which are interpreted and analyzed in this section.

Finally, the last chapter is the discussion chapter which contains thoughts about future work or what could be changed if this project was done again.

2 Background

The background chapter of this thesis will give a basic explanation of how non-life insurance works as well as theoretical background to the mathematical tools that are used to analyse the given claim data with.

Even though some of the background is about the behaviour of extremal events, the scope of this thesis is not to model actual extremal events but rather to find the limit for where the claim data can be considered too extreme to fit the distributions that model the regular claim data.

2.1 Non-life insurance

Non-life insurance is in essence the transfer of risk of certain unpredictable events that incur financial losses, from individuals to an insurance company. The freedom of risk for the individual is bought for a price, a *premium* that is paid regularly for as long as the risk belongs to the insurance company.

The events are pre-specified and if they occur the insurance company covers most or all of the financial loss for the customer.

To then set the premium for a specific individual, information about her is utilized to provide an as accurate approximation as possible for her expected cost. For example, a person's residence area and age are common factors of consideration. Having individual prices is essential since over- and underpricing the premium in relation to the true expected cost of the customer are both potential monetary losses for the insurance company. In the case of underpricing, the financial losses of the customers will be greater than the premium, and in the case of over-pricing, there is an increased risk of customers changing to insurance companies with a cheaper premium. Thus, you want to obtain an as accurate premium as possible that still gives a net profit for the insurance company.

However, when using the historical data to determine the premium price, one thing to watch out for is outliers. An outlier in this case could be a person who suffers a great financial loss due to an incredibly rare event that can be assumed to be independent of that specific person's attributes such as area and age.

If this great financial loss, *a large claim,* would be treated as normal, and thus used to model the premium for all those of the same area and age, an unnecessarily large premium would be set for those customers. This would overprice the premium of people with that age or area since that rare event should rather be modelled as independent of that specific group of customers and instead be modelled as a financial loss of the entire custom field. The result of dealing with large claims in this matter is that the premium for everyone increases slightly rather than one group having to cover for all of it.

2.2 General

Amongst the most common models to analyze the insurance claim data are the lognormal and the Pareto distributions. But the two distributions cannot by themselves fit the entirety of the data. The lognormal distribution does not describe the right tail of the data well enough since it is often much more positively skewed, with very large claims appearing more often and with higher magnitude than the lognormal predicts. It does however fit well at the lower spectrum where there are many claims of low magnitude (Ananda & Cooray. 2007)

On the other hand, the Generalized Pareto distribution (GPD) captures the heavy-tailed behavior of the data with much better accuracy. However, this distribution lacks in its description of the right tail, as it does not provide nearly enough of the high frequency, low cost claims.

These two distributions can, given a certain threshold that divides the data, complement each other to describe one part each. If a threshold is attained, so that the GPD has as good of a fit to the excess of it as possible, then the part below the threshold can better be modelled with log-normal type distributions. The objective to find the large claim limit can then be simplified into finding the best fit for the GPD distribution, which includes the threshold limit.

2.3 Generalized Pareto distribution

When describing the outcome of an insurance portfolio, it is of great importance to include the occurrence of extremal events. These rare occasions can due to their mere size have a detrimental effect on the pricing for the insurance holders. Specifically, small groups can receive overpriced premiums if a large claim affects a person within it.

One of the most useful methods of describing extremal events is fitting the data with a GPD, the *Generalized Pareto Distribution* to the upper tail. (which is defined as the data above a certain threshold). (Embrechts, et al. 2012)

Here the thought is that for the threshold that the GPD distribution has its best fit, the data below the threshold is considered non-extreme and can thus be treated regularly.

Let *u* denote the threshold and *X* be a random variable with distribution function F. The excess function over the threshold *u* is now given by

$$F_u(x) = P(X - u \leq x | X > u\ ),\ x > 0. \tag{2.1}$$

And the generalized Pareto distribution is given by

$$G_{\xi,\sigma}(x) = 1 - \left(1 + \frac{\xi x}{\sigma}\right)^{-\frac{1}{\xi}}, \quad \xi, \sigma, x > 0. \tag{2.2}$$

Where σ is the scale parameter and ξ is the shape parameter (Hult, et al. 2012 and Davidsson & smith. 1990).

2.4 Finding the threshold u

Computing the threshold *u* is to balance bias and variance. From a high threshold follows a larger variance due to smaller sample size. However, it also makes the underlying GPD approximation more accurate since it scopes in more on the right tail, where the fit is at its best.

2.4.1 Mean residual life plot

Also called the mean excess function, this is a graphical method of determining the threshold limit *u,* that is based on the mean value of the GPD that is assumed above the threshold limit (Embrechts, et al. 2012).

First let $X_1, \dots, X_n$ be a series that describes the excess values of the threshold u_0. Based on the expected value of GPD

$$E(Z) = \frac{\sigma}{1-\xi}, \quad \xi < 1 \tag{2.3}$$

you get the following equation:

$$E(X - u_0 | X > u_0) = \frac{\sigma_{u0}}{1-\xi}. \tag{2.4}$$

Here, ξ still has to be less than 1 and σ_{u0} is the GPD scale parameter. u_0 is the threshold.

By the threshold stability property of the GPD, if the distribution is valid for exceedances over the threshold u_0, then it is also a valid model for the exceedances of all thresholds larger than u_0. Thus, for $u > u_0$:

$$E(X - u | X > u) = \frac{\sigma_u}{1-\xi} = \frac{\sigma_{u0} + \xi * u}{1-\xi} \tag{2.5}$$

The left-hand side of the equation above is the mean excess of the threshold. This can be estimated by the sample mean of the excesses over the thresholds u. This is done for different values of u, ranging from u_0 to just below the largest sample value (Embrechts, et al. 2012 and Coles. 2001).

Expressed mathematically:

$$\left\{ (u, \frac{1}{n_u} \sum_{i=1}^{n_u} (x_{(i)} - u)); u < x_{max} \right\} \tag{2.6}$$

Here, $X_1, \dots, X_n$ represents the exceeding observations. The result of this is a locus of points with the threshold as x-value and the mean exceedance as the y-value. These points constitute the mean residual life plot. In this plot, a value of u where linearity is attained after is looked for. (Embrechts, et al. 2012).

By construction, toward the right side of the plot where the threshold gets very high, only a few observations will exceed u and thus create instability. This is something to consider when interpreting the plot. Worth mentioning is also the general downside of a graphical interpretation method, namely inaccuracy and subjectivity.

2.4.2 Parameter Stability plot

The parameter stability plot is another graphical method to determine the threshold u. Similarly to the mean residual life plot, the fact that if exceedances of u_0 is correctly described by a GPD, then exceedances over a threshold higher than u_0 also follow a GPD. In addition, information about the parameters is also kept. If the GPD over u are ξ and σ_{u0}, then for any threshold larger than u_0, the parameters are $\xi_u = \xi$ and

$$\sigma_u = \sigma_{u0} + \xi(u - u_0) \tag{2.7}$$

Now, to remove the dependence of u let:

$$\sigma^* = \sigma_u - \xi_u * u \tag{2.8}$$

Now $\sigma^*$and ξ_u are plotted against u. u_0 is the suitable threshold when σ^* and ξ_u remains constant for all $u > u_0$ (Coles. 2001).

2.4.3 Rules of thumb

Yet another method to determine *u* is via the rules of thumb.

Suppose the data is an ordered sequence $X_1, \dots, X_n$. The threshold is basically the *k*:th upper order statistic $X_{(n-k+1)}$, also called the tail fraction. As *n* tends to infinity, so does *k*. To ensure tail convergence, $k/n \rightarrow 0$ as $k, n \rightarrow \infty$, i.e. *k* has to grow slower than *n*. This condition ensures that as the sample size grows, so does the quantile level but with a faster rate (Mcdonald & Scarrot 2012).

With this condition satisfied, the following threshold methods have been derived (Loretan & Philips. 1994 and Ferreira, et al. 2003)

$$k_1 = \sqrt{n} \tag{2.9}$$

and

$$k_2 = \frac{n^{2/3}}{\log(\log(n)} \tag{2.10}$$

2.5 Model fit

When fitting a GPD to the excess of a threshold, both the probability-probability and quantile-quantile plots are good tools when evaluating the goodness of the fit. (Coles. 2001).

The quantile-quantile (qq) plot is a graphical method that can be used to determine how well a data set belongs to a given probability distribution. It plots the quantiles of the data and the attempted fitted distribution against each other, and if the fit is good the graph will take shape of a straight line.

Probability- probability plot is also a graphical method, but unlike the qq-plot, the cumulative distribution functions (cdf) are plotted against each other. A good result is here also indicated by straight line in the plot.

2.6 Generalized Linear Model

The generalized linear model (GLM) is generalization of linear models. Ordinary linear models compute the expected value of a given response variable as a linear combination of predictors. This means that that type of model is a linear-response model, i.e. a constant change in the predictor, *X*, gives a constant change in the response variable *Y*.

One can intuitively see that this is an inappropriate quality for response variables without a normal distribution behavior. Instead, in a GLM, the response variable is assumed to come from a distribution in the exponential family. In addition, in contrast to a regular linear model, the response variable must not vary linearly but rather an arbitrary function of it. varies linearly. This function is called *g,* the *link function.*

Expressed mathematically, the expected value of the response variable is as follows:

$$E(Y) = g^{-1}(X\beta) \tag{2.11}$$

In this thesis it was suggested by Trygg-Hansa that the multiplicative version of the GLM was used. In that case, the response variable *Y* can be modeled as

$$ln(u_{ij}) = ln(y_0) + ln(y_{1i}) + ln(y_{2j}) \tag{2.12}$$

In the case of two tariff variables, where i and j are the i:th (and j:th) group of the tariff (predictor) variable. For example, if the first tariff variable is age then index 1i refers to the i:th age group, perhaps 25-45. As all of the parameters y are estimated, the regression can then predict values of y^* with

$$y^* = y_0\, y_{1i} y_{2j} \tag{2.13}$$

(Johansson & Ohlsson, 2010).

2.6.1 Parameter estimation (MLE)

The parameters of a GPD can be calculated in different ways, Maximum Likelihood is usually preferred. Let $y_1, \dots, y_n$ be n excesses over the threshold u. By using that the likelihood function is

$$L(\xi, \sigma) = \prod_{k=1}^{n} g_{\xi,\sigma}(y_k) \tag{2.14}$$

Where

$$g_{\xi,\sigma}(y_k) = \frac{1}{\sigma}\left(1 + \frac{\xi y_k}{\sigma}\right)^{-\frac{1}{\xi}-1} \tag{2.15}$$

This gives

$$logL(\xi, \sigma) = -nln(\sigma) - \left(\frac{1}{\xi} + 1\right)\sum_{k=1}^{n} \log\left(1 + \frac{\xi}{\sigma} y_k\right) \tag{2.16}$$

Which is called the log-likelihood function (Hult, et al. 2012).

2.7 Champions' model

The Champions' model is a graphical method used to compare the performance of two models against a reference data set. The method was suggested from Trygg-Hansa.

First, two samples are created from the original data set. One that will be used to obtain the model parameters for the models to be compared and the other as reference to compare to the models to. These two data sets are called the model set and reference set respectively. The model set will consist of approximately 80% of the original data while the reference set the remainder.

The next step is to divide the model data into groups. For instance, let one group consist of data with a certain range of the prediction variables, let's say ages between 25-30 and duration between 0-10 years. When the whole population is covered by non-overlapping groups, obtain

the value that you want to measure your model performance with trough GLM. These different groups will be the dummy variables in the GLM and the first group of each variable is set as the intercept. In this thesis the response variable in the GLM is the *burning cost.* This is the cost per exposure for Trygg-Hansa for given profile. Let that value be called $Y_{i,j}$ and the true burning cost from the reference data T_i , where

$$i = 1{:}\,n$$
$$j = 1{:}\,2$$

$i, j$the data is collected from, respectively.collected from, respectively.

Compute the $Y_{i,1}/Y_{i,2}$ quotient across all groups (for every i). In addition, gather the reference value for the burning cost from the reference data set across all groups. The result from these processes can be gathered in a table as follows:

	Group 1	**Group 2**	**Group n**
Burning cost, model 1	$Y_{1,1}$	$Y_{2,1}$	$Y_{n,1}$
Burning cost, model 2	$Y_{1,2}$	$Y_{2,2}$	$Y_{n,2}$
True Burning cost	T_1	T_2	T_n
Quotient	$Y_{1,1}/Y_{1,2}$	$Y_{2,1}/Y_{2,2}$	$Y_{n,1}/Y_{n,2}$

Table 2: *Table of the Burning cost and quotient*

The next step is to lump groups with similar quotients together, taking the mean of their values. The quotients are lumped together by taking for example the lowest 10% of all quotients in one point, and then the next to lowest 10% in the next. Finally, a plot is made where the horizontal axis is the lumped quotients and the vertical axis is the mean value of the different burning costs. This is computed cost for a group per year of exposure. The graphical interpretation takes place by finding out which model fits the true burning cost, the best. In order to make the interpretation easier, the models are scaled so that

$$\sum_{i=1}^{n} Y_{i,1} = \sum_{i=1}^{n} Y_{i,2} = \sum_{i=1}^{n} T_i. \qquad (2.17)$$

Now, the total claim size for the three plots are all equal and the interpretation of which model fits the true data the best can be done more easily.

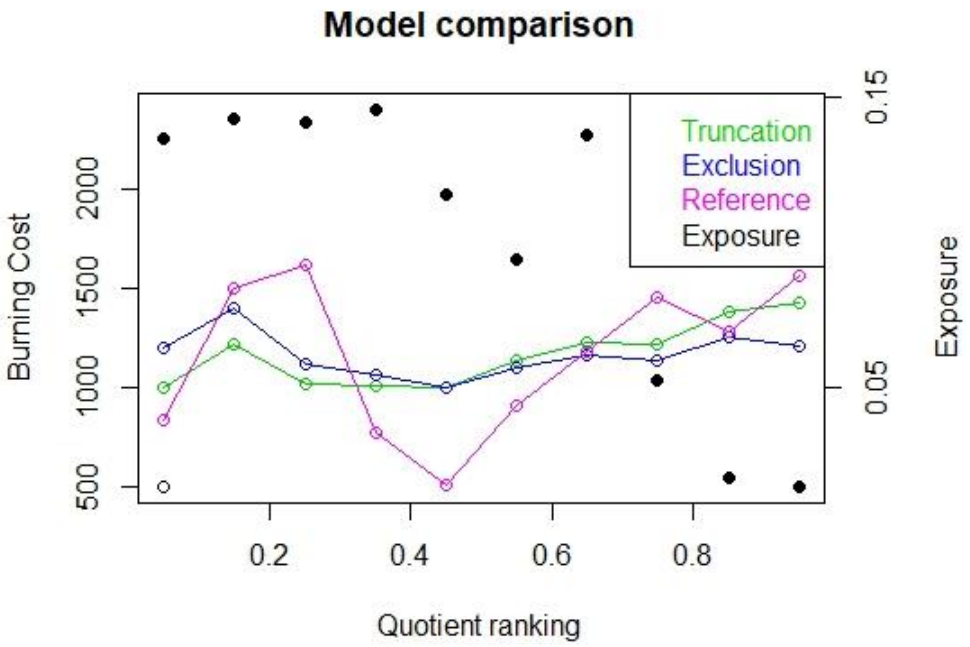

Figure 1: *Example plot of the Champions model*

For every lumped quotient, the total amount observations (total exposure in the case of this thesis) in those groups are computed and added to the point. This is done to quickly determining the reliability of that specific point in the plot. In the plot above, these are the black dots.

2.8 Data set

The Data set that was used was manipulated data from Trygg-Hansa. Due to discretion it has been decided to not include the actual data in the thesis. A description of the data is as follows:

The data consists of 20 columns and over 5 million rows.
Each row in the data set can be seen as one insurance holder over a period of time where each row contains information about his/her profile. Observe that one insurance holder may appear on several rows depending on renewal of the insurance terms or the type and/or number of damages. The columns that were used contained the following information about the insurance holder:

Column name	Description
skar	The year when the damage occurred.
Produkttypkod	This is referred to as the product code, i.e. what type of coverage the damage belongs to. See all three types of coverage below.
Produkttypkod-RP	This is extended traveling coverage, for instance if you need medical help abroad.

Produkttypkod-BY	This is building coverage, for example the result of water leakage.
Produkttypkod-PL	This is base coverage, for instance furniture in the building.
Aalde2	Age of the insurance holder
Durat6	The duration insurance holder has been insured by the same company.
Byald	Age of the building.
Boyta	Size of the building.
Antper	Number of insured people.
Bebygg	The type of building.
Riskar	Amount of years that the risk belonged to the insurance company. This is the measurement of exposure.
Ultimo	The total cost of the damage.
skadenr	Damage ID.

Table 1: *Description of data*

The most important information in this table is the Product codes. The Product code denotes the type of coverage that the claim belongs to and thereby each code can have their own underlying distribution since different types of coverages have very different expected costs and variances. For example, given that a claim is of the BY coverage (building coverage) product code, the expected claim size can increase tenfold compared to that of a PL (base coverage) type product code. Because of this, the data is divided and analyzed for each product code separately, and a threshold limit is obtained for each one.

3 Execution & result

In this chapter there are two sections. In the first section the methods used to obtain the results are provided. The second chapter is constituted by the results and their interpretation.

3.1 Execution

The process of treating the data set and implementing the methods on it is here described and can be divided into several steps.

3.1.1 Modify data set

The data set will be modified in the following two ways:

3.1.1.1 Merge insurance claims

In the data set received, each row does not necessarily capture the entirety of a specific damage for the insurance holder. One damage or insurance holder may appear on several rows in the set.
For example, one insurance claim on 1 million kronor (a large claim) might be divided into four rows in the data set. Without modifying the data, it will appear as four independent smaller insurance claims together adding up to 1 million kronor. With the help of the damage ID, these different rows are added together so that the insurance claim is correctly counted as a large claim.

3.1.1.2 Divide product code

As earlier described the data set consist of three different product types (RP, BY and PL).
As the value of these types of coverages can vary quite a bit the data is operated on these three types separately. This means that there will be one large claim limit for each product type.

3.1.2 Large Claim threshold

To find the large claim limit, the graphical methods that are the mean residual life plot and the parameter stability plot, are used. “Rules of thumb” is also used to find the large claim limit.

With maximum likelihood estimation, a GPD is fitted to the data exceeding this limit and the goodness of this fit is analyzed and measured with the help of qq and pp-plots. The limit that gives the best fit of the GPD is considered to be the optimal large claim limit.

3.1.3 Champion model

Once estimations of the large claim limit have been obtained through the previous methods, the Champions’ model is utilized to determine whether truncation or exclusion above the threshold limit should be used.

The steps to perform this method can now be divided into five steps.

3.1.3.1 Sampling

First, two samples are created, taking random chunks from the original data. One large sample of 80% and one smaller of the remaining 20%. The larger sample is the model data set and the other is called the reference data set.

3.1.3.2 Finding suitable groups

When choosing how to divide the data, one must make sure that the groups do not become too small in order to have reliable values of that group in the reference data set. Before the groups where chosen, plots of the different predictors were made against the burning cost, to find patterns of how the groups should be split to decrease the variance within them. The groups must not overlap since each data point should only occur once in the data set for the GLM to function properly. An example of a group on an analysis with three variables (age, duration, people) would be every data point in the set that meet the following conditions: Has age 16-26, duration below 5 years and up to three people on the insurance. One of these groups is defined as the intercept group for the GLM analysis.

For different product codes, these variables, or *predictors,* affect the burning cost differently. Thus, one set of plots was done for each product code.

For RP and BY plots of the burning cost were made against Age of insurer, duration of insurance, number of people on the insurance and finally building size. For RP, the travel insurance, predicting variables regarding building properties were excluded and thus the plots were only for the other three predictors named above.

A plot of this type may look like this:

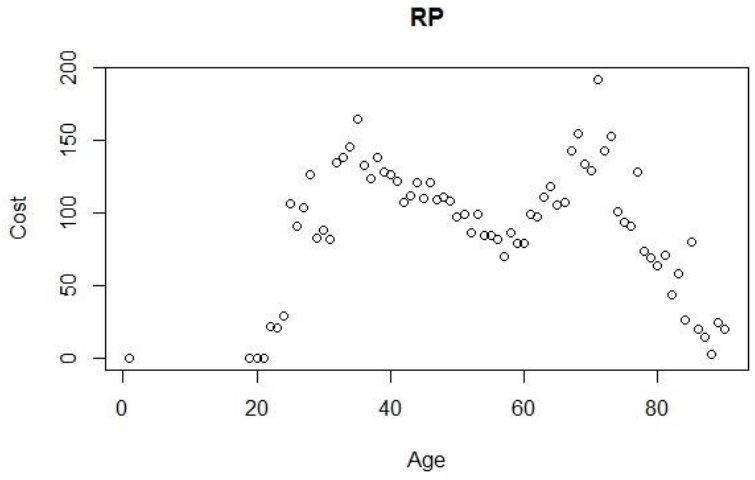

Figure 2: *Example plot*

Choosing groups with this plot in mind would mean trying to avoid a large gap in mean value within any given age group. For example, one could argue that there is a clear mean value difference between the ages of 20 and 40. Thus an age group of 20-40 should be avoided and

instead split into perhaps 18 -26 and 27-40. This should be done for all product types and each of their predicting variables.

3.1.3.3 Obtain burning cost

The estimated burning cost is computed for the truncation model and the exclusion model by estimating values for all different groups with the parameters achieved by the GLM. On the reference data set, the true burning cost is calculated raw.

With a value obtained for both models for each group, the total burning cost for each model was multiplied with a factor to bring the net burning cost for each model to the same magnitude as the reference data. This makes the graphical interpretation easier.

3.1.3.4 Plotting

For each product code with its calculated threshold limit, the estimated burning cost of the two models are plotted together with the reference data set's true value of the burning cost.

3.1.3.5 Graphical interpretation

In each plot, the performance is measured by comparing which of the models Truncation and Exclusion that are closest to the reference data burning cost. The closer to the true value, the better the model resembles reality. In points of very little exposure, any deduction is taken with a grain of salt since those numbers might be very unreliable.

3.2 Result

The final result of the thesis consists of two parts, one where the threshold limit is attained for each product type and the second part where it is determined whether truncation or exclusion above the threshold limit is optimal. Including in the result section however, are the sub results: What groups that were chosen for the GLM analysis and its estimated parameters.

3.2.1 Large claim limit

The graphical methods have their drawback of being subjective and not exact. However, if several graphical methods all point in the same direction, a rough estimation of the threshold can be obtained.

In the mean residual life the threshold is chosen where the curve becomes linear and for the parameter stability plot it is chosen as the value where the plot is constant afterwards. The analysis is done on the data for each product code separately.

From the graphically obtained limits, the two best candidates are used and a GPD is fitted to the exceeding data for each.

3.2.1.1 RP

For the product code RP, the following plots for the graphical methods were achieved:

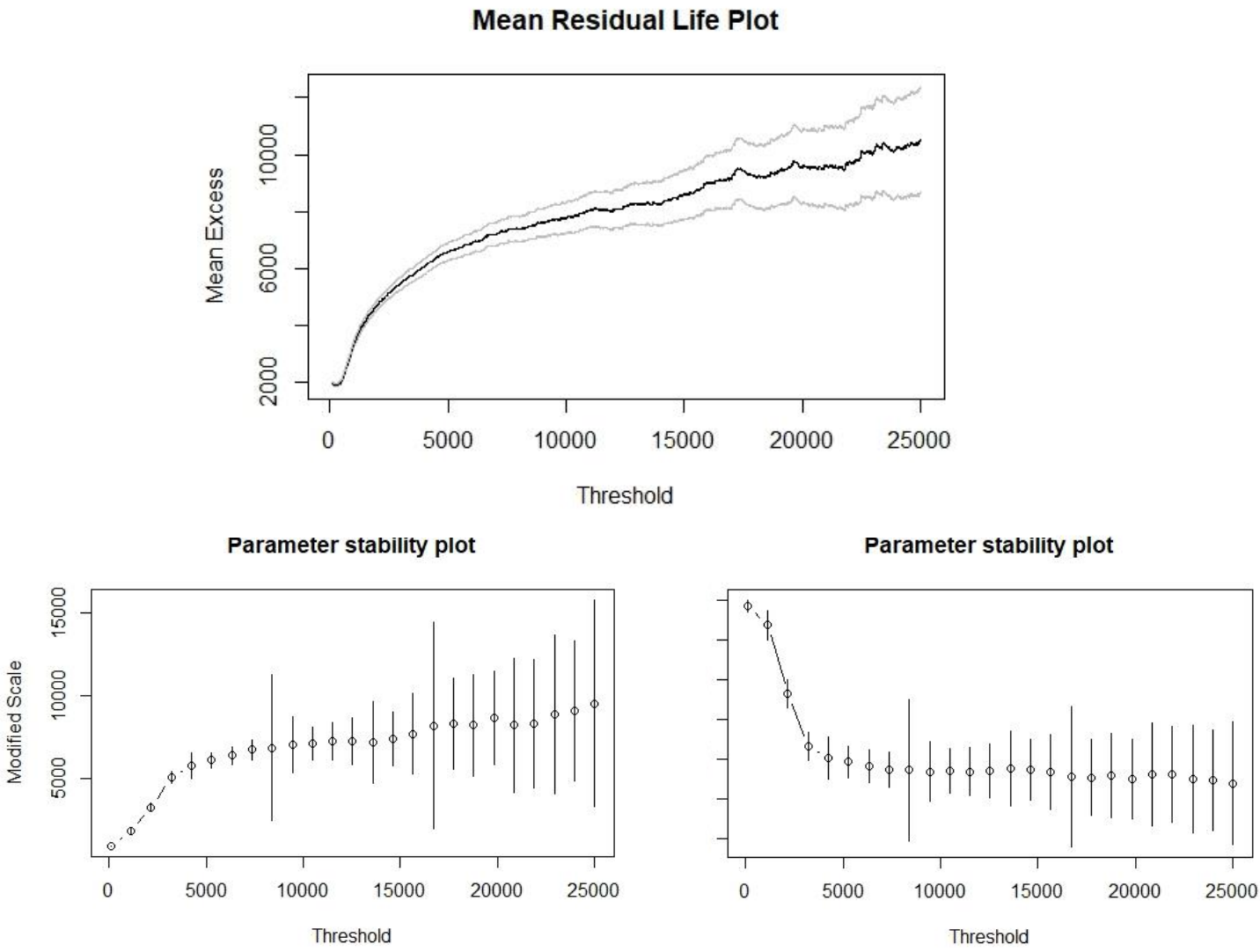

Figure 3: *Graphical plots for the large claim limit for the product code RP*

These plots point at a limit of around 8 000. This value is taken from where the MRL-graph appears to become linear and where the parameter stability plot appears to become constant. Additionally, the rules of thumb method achieved the following two threshold limits:

Method	k_1	k_2
Value	24400	17040

Table 3: *Rules of thumb result*

For each limit, a GPD is fitted on the exceeding data. The following QQ-plots are, in rising order of the limit (8 000, 17 000 and 24 000 respectively) what determines the goodness of the fit.

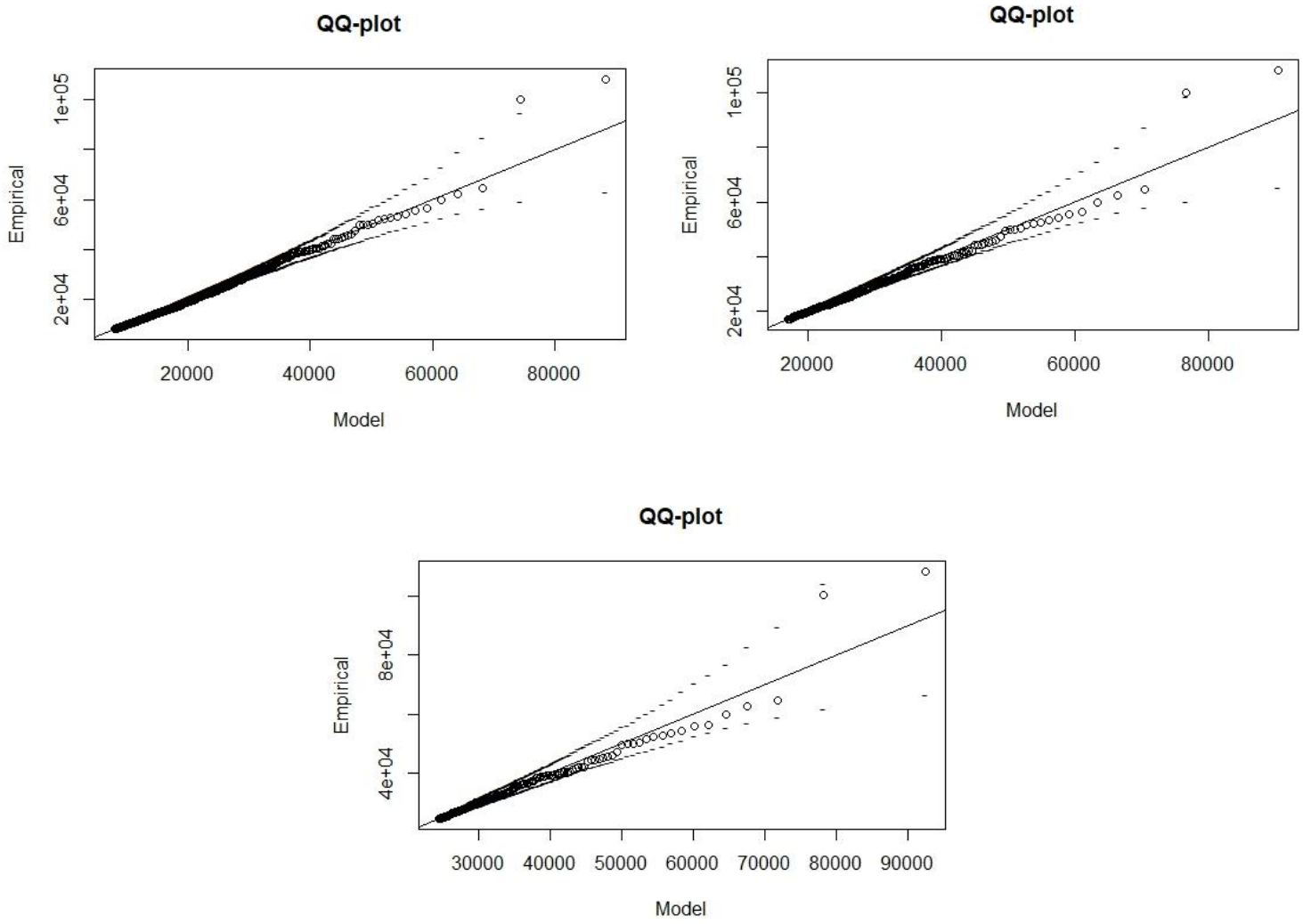

Figure 4: *QQ-plots for the fitted GPD:s*

It is difficult to exclude any limit as they all look quite similar. Thus, to choose a limit, the PP-plots are looked at. These are in the same order as above.

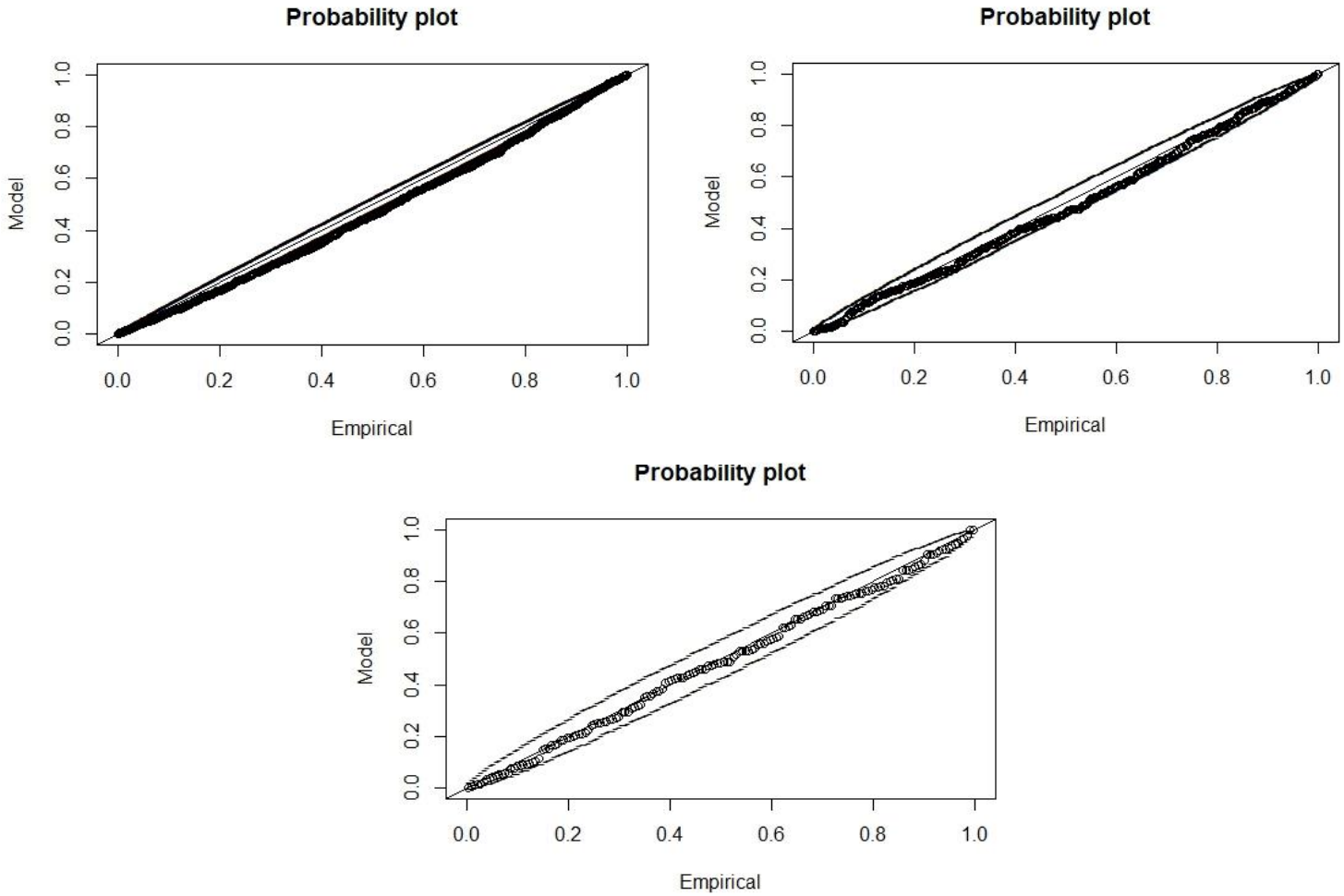

Figure 5: *Probability plots for the fitted GPD:s*

In the PP-plots the best fit seems to be by the last limit of 24 000.

3.2.1.2 BY

For the product code BY, the following plots for the graphical methods were achieved:

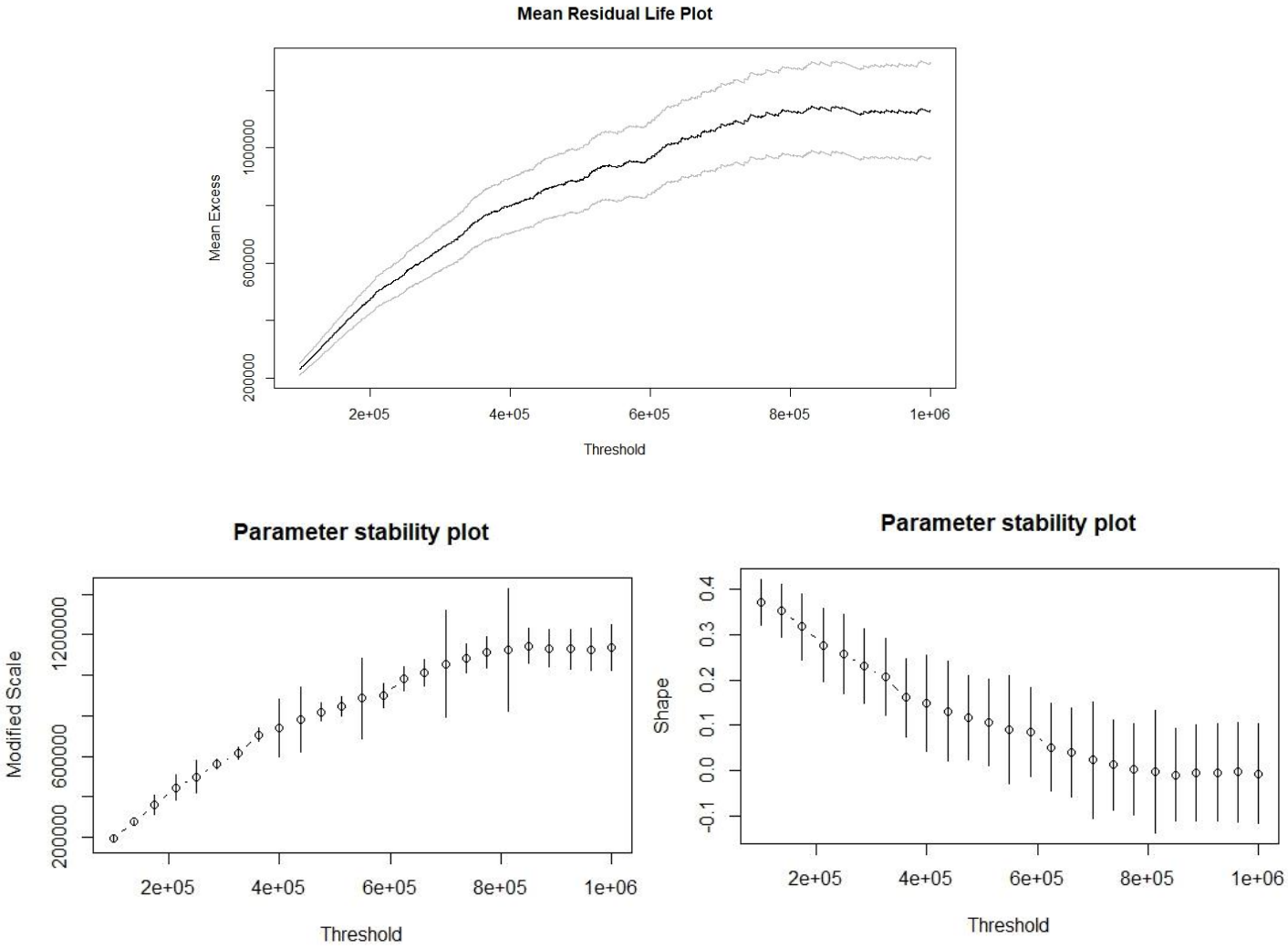

Figure 6: *Graphical plots for the large claim limit for the product code BY*

As described in the method, the threshold limit is indicated at the point where the graphs become linear. In this case both the parameter stability plot and the mean residual life plot point in the same direction. 800 000 is taken as the threshold limit for this product code.

In addition to the plots, *rules of thumb* achieved the following limits:

Method	k_1	k_2
Value	606 020	248 160

Table 4: *Rules of thumb result*

As the result vary, QQ- and PP-plots are used to further determine which of the obtained thresholds provide the best distribution fit to the upper spectrum of the data. Following are the QQ- and PP-plots with the computed threshold limits (in rising order):

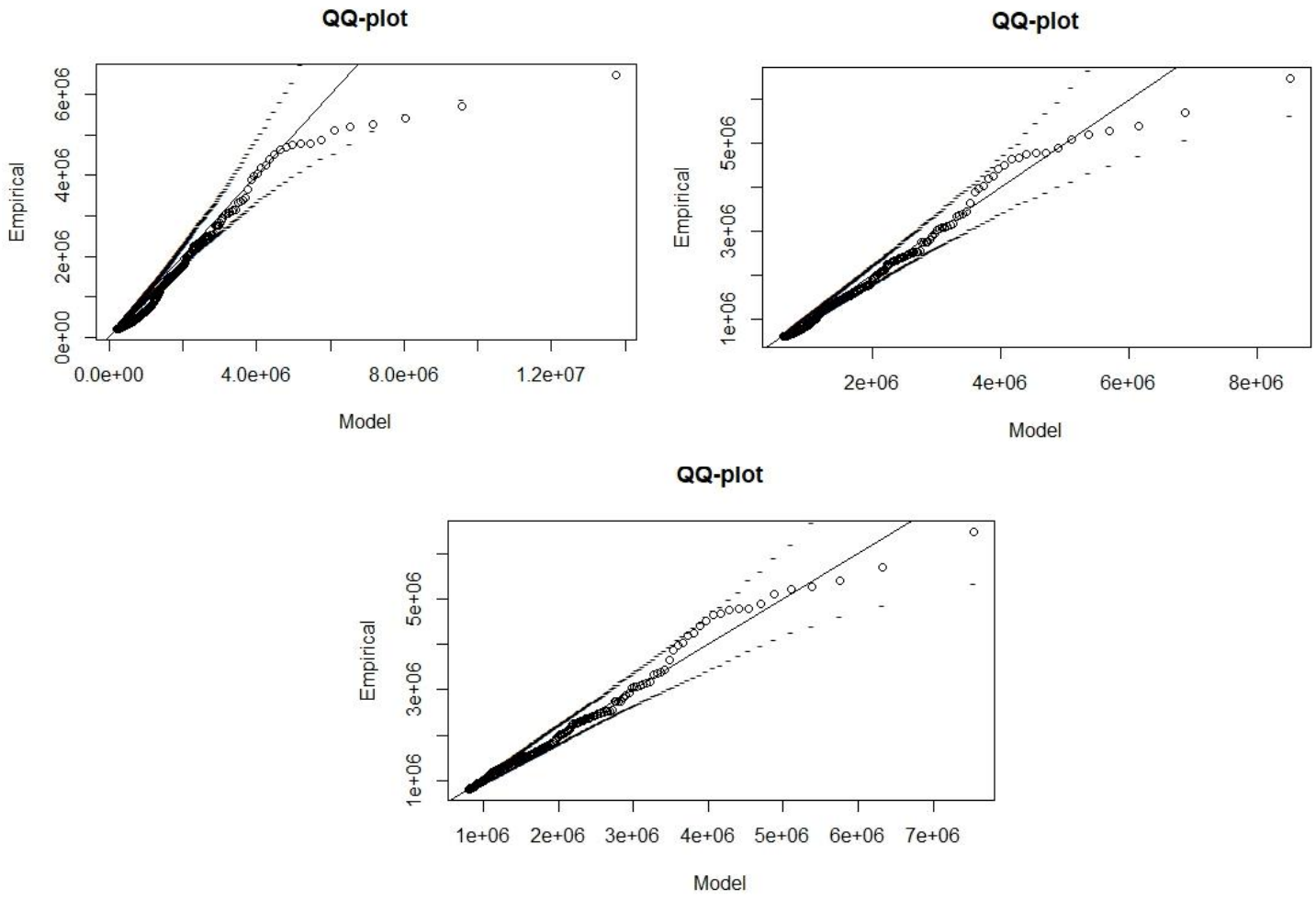

Figure 7: *QQ-plots for the fitted GPD:s*

Above are the QQ-plots for the limits 248 000, 606000 and 800 000 respectively. The best candidates for the threshold limit are the two last plots, but further evaluation is needed to exclude either of them.

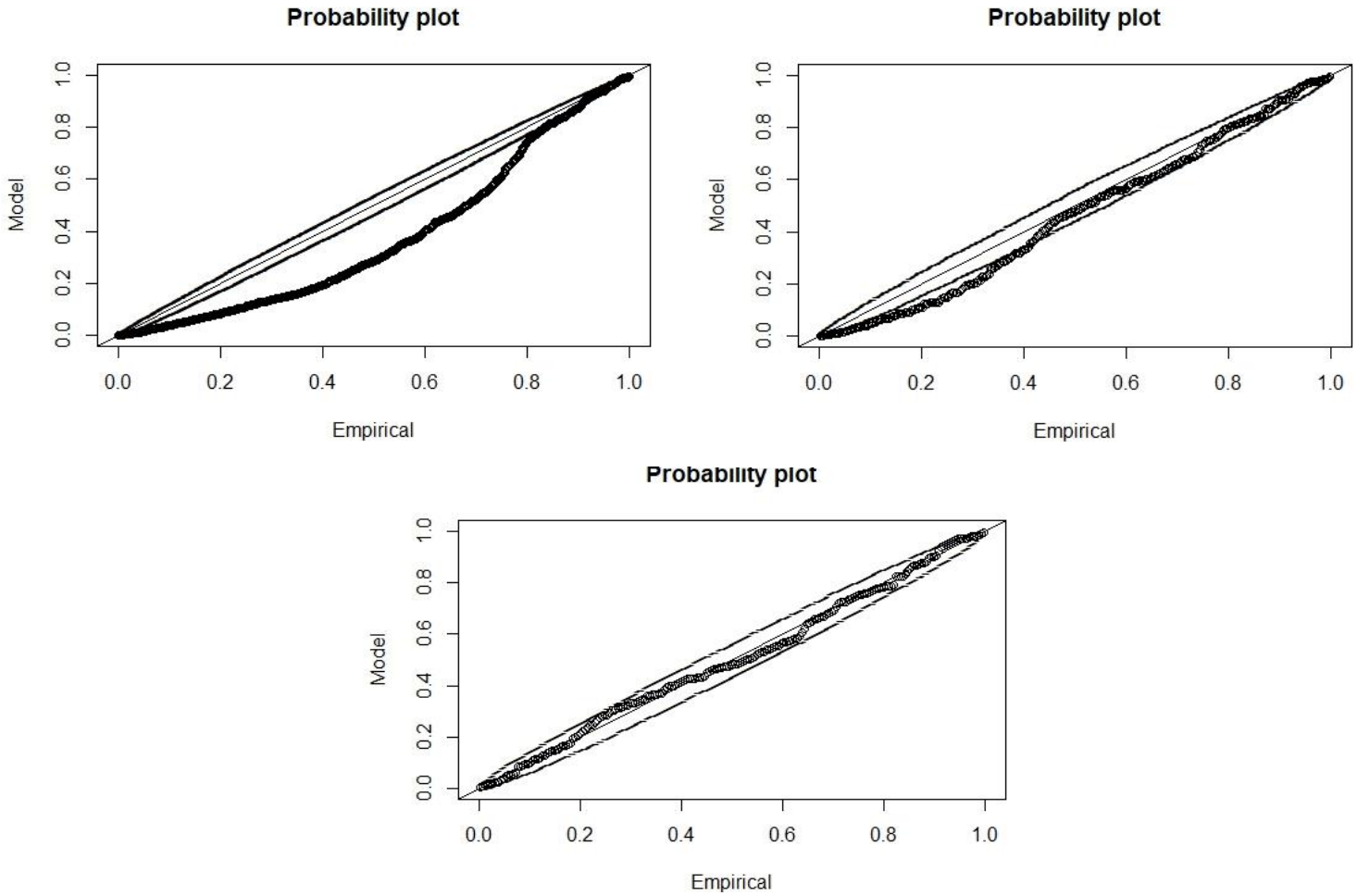

Figure 8: *Probability plots for the fitted GPD:s*

The pp-plots for the limits in same order as the QQ-plots. The best linearity is achieved in the last plot; thus the 600 000 limit is finally excluded and the threshold limit for the BY product code is chosen: 800 000.

3.2.1.3 PL

For the product code PL, the following plots for the graphical methods were achieved:

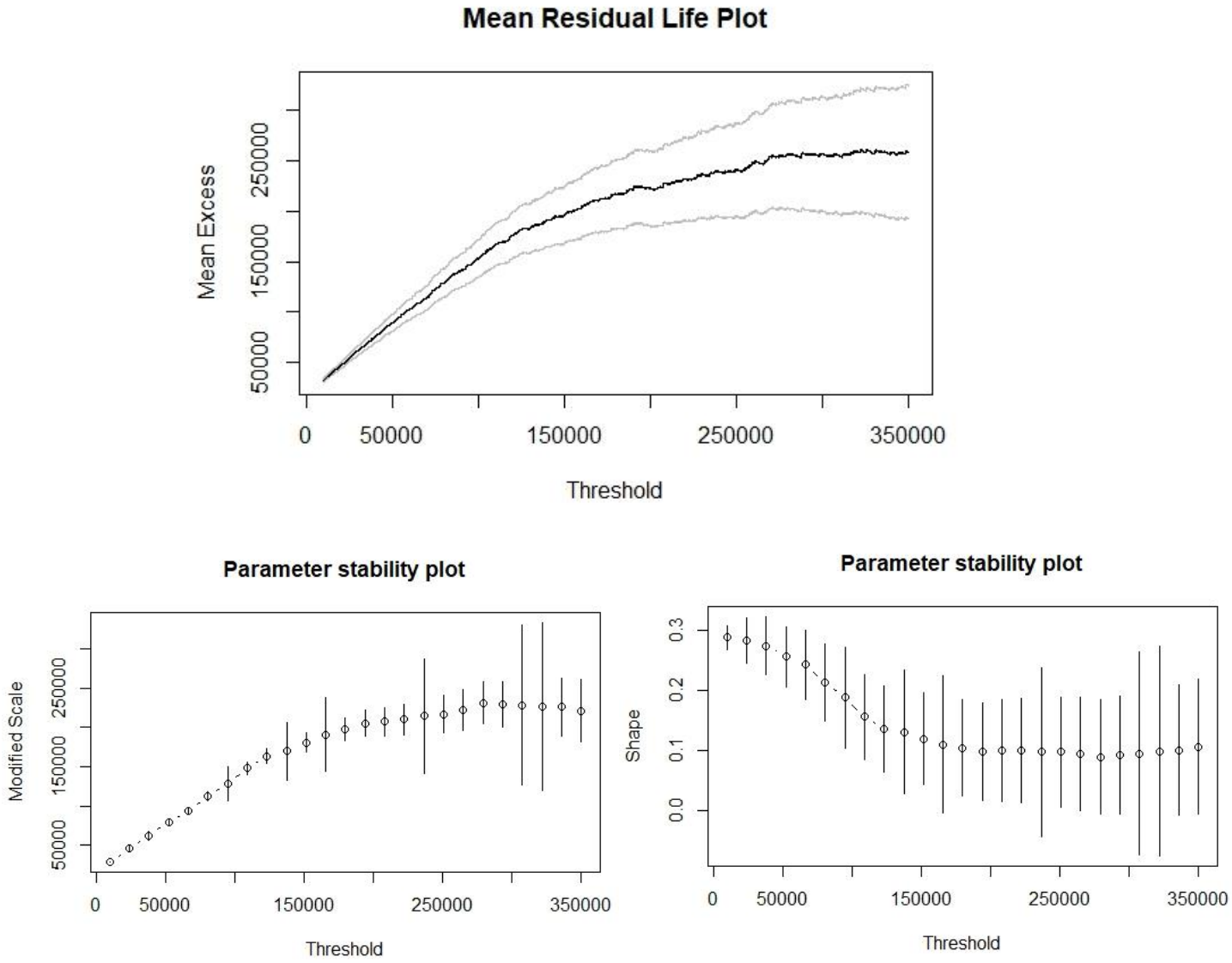

Figure 9: *Graphical plot for the large claim limit for product code PL*

As can be seen, the threshold limit is very different to the one of BY. It is much lower and looks to be somewhere around 275 000. This is expected since the data set of PL contains much lower claims.

The rules of thumb achieved the following threshold limits:

Method	k_1	k_2
Value	212 324	98 673

Table 5: *Rules of thumb result*

Below are the QQ-plots for the achieved threshold limits of PL, in rising order.

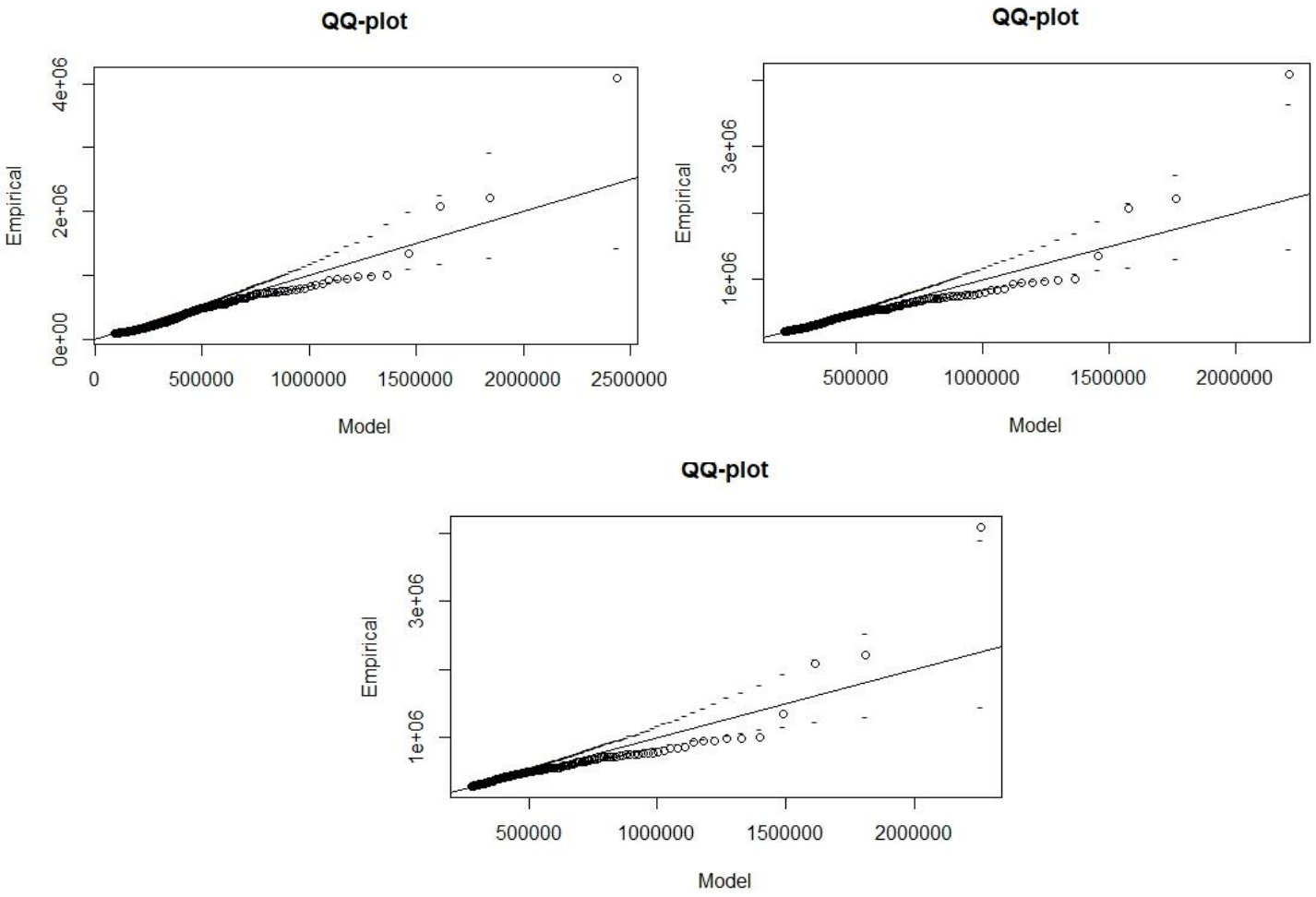

Figure 10: *QQ-plots for the fitted GPD:s*

The QQ-plots hardly give any reason to exclude either of the limits. However, if you examine the PP-plots:

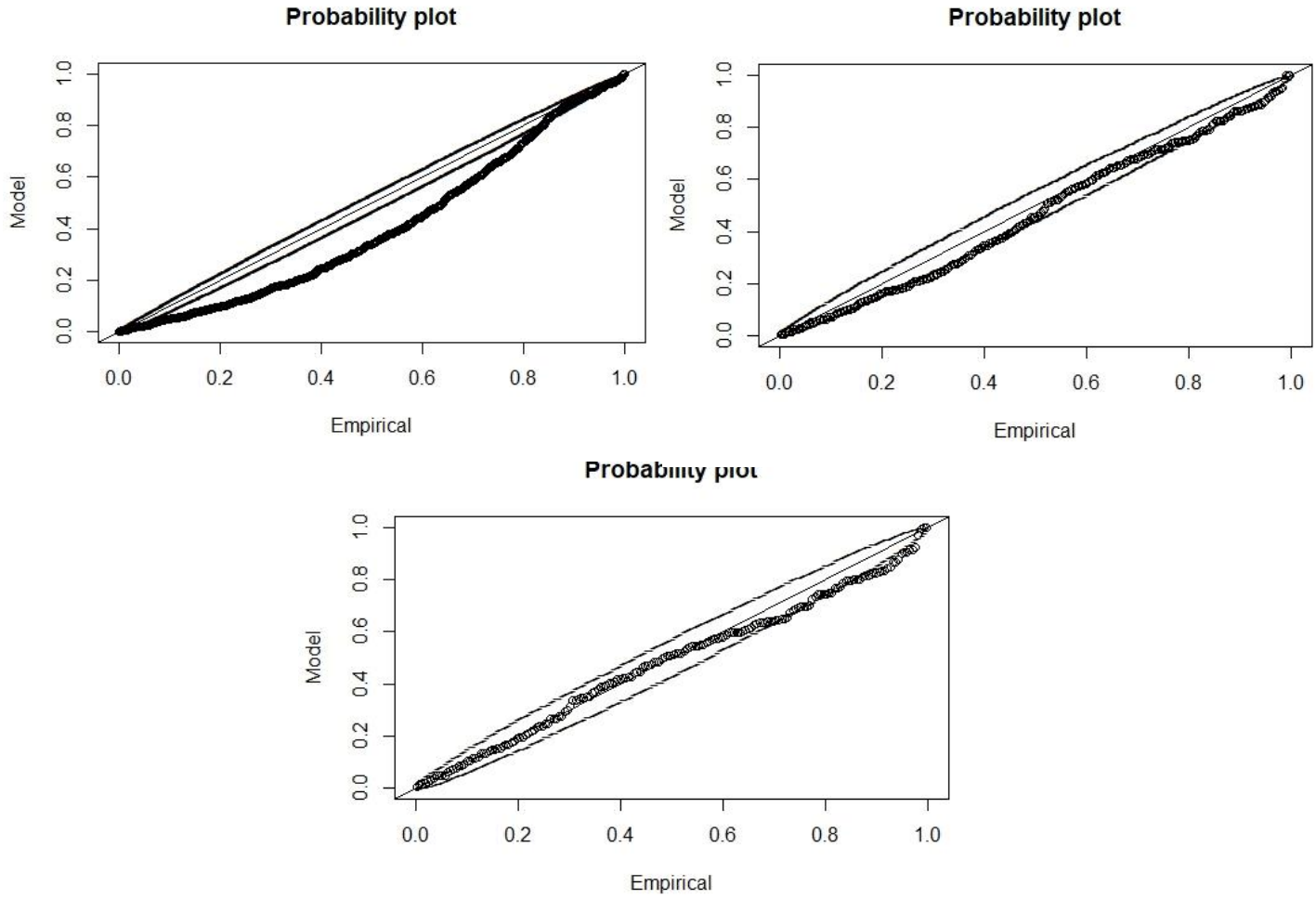

Figure 11: *Probability plots for the fitted GPD:s*

The last plot with the limit of 275 000 provides the best fit. Once again, the graphical methods have been superior to the rules of thumb.

3.2.2 GLM analysis

The groups of predicting variables that were used in the champion model had to be chosen wisely. On the one hand, they had to be large enough so that the reference data of that group would have reliable values.

On the other hand, if the groups were too large, interesting connections between a varying predicting variable and the response variable could be missed.

Three to four predictor variables were used at a time. To choose the groups for these variables, the response variable (burning cost) was plotted against the predict variables, one at a time and for each product code.

The groups were chosen by visual inspection of this plot so that the gaps in the y-axis would not be too large. Below are the plots that were used to choose the groups. For each product code there is also a table with full detail of the groups as well as the estimated parameters from the

GLM analysis, one for exclusion and one for truncation. These values are later used in the Champions model to represent the two compared models.

3.2.2.1 RP

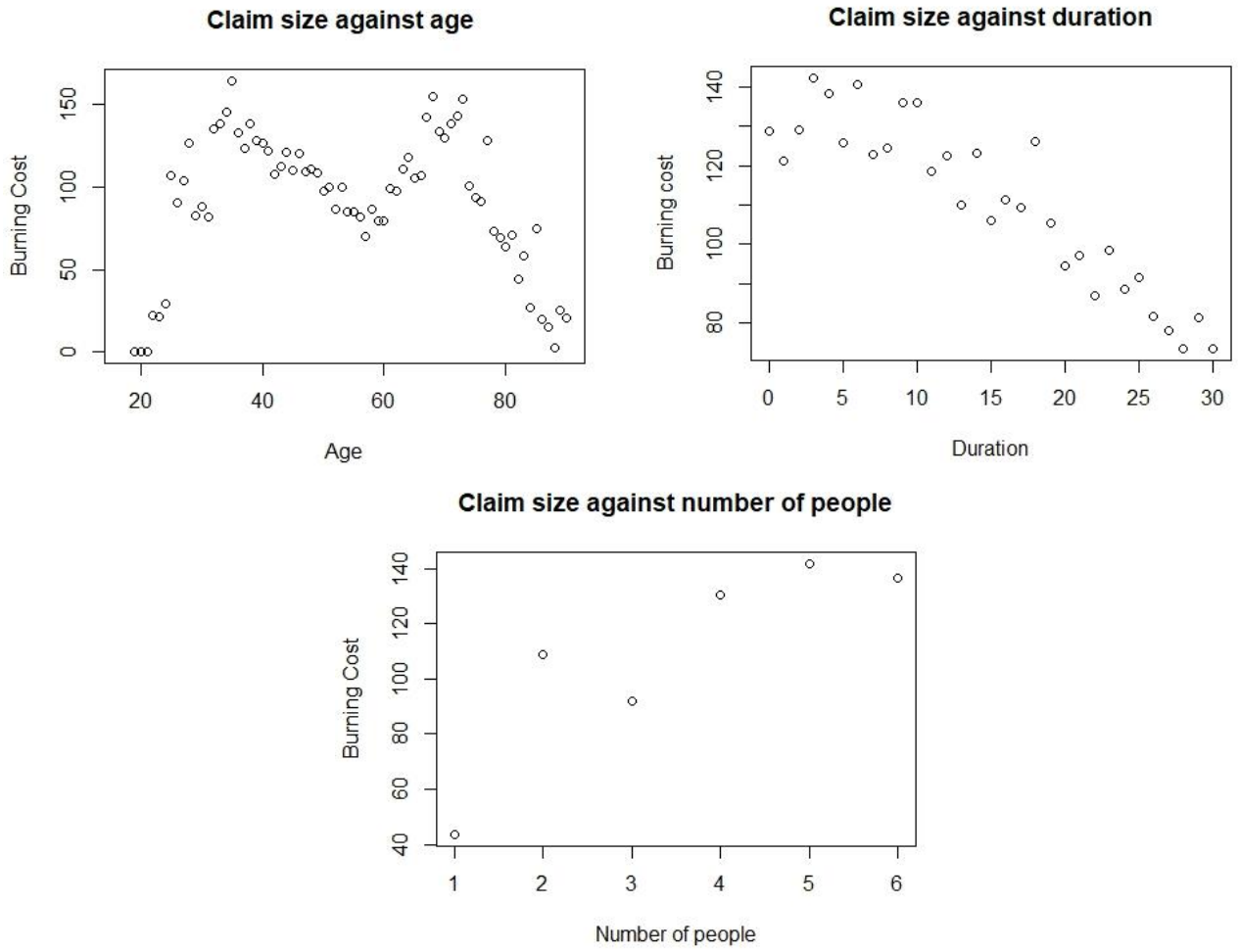

Figure 13: *Plots of burning cost against various predictor variables*

Here, the response variable (burning cost) was plotted against the age, insurance duration and number of people on the insurance (respectively). For product code RP, information about the insurer's building is hardly relevant and was chosen to not be plotted. By visual inspection of the above graph, groups were chosen. The train of thought was to try and separate clear shifts in mean value into different groups. For instance, between the age of 30 and around 40 there is a clear increase of the burning cost. Two age groups to start with would then be 16-30 and 31-48. This train of thought was applied for all of the graphs.

The result for RP was five age groups, four duration groups and three (amount of) people groups. The exact ranges in every group can be seen in the table below. The intercept in the GLM is the group you get if you take the first sub group from each category. The result for the GLM analysis for exclusion and truncation can also be seen in the table below:

	Tariff variable 1, Age		Tariff variable 2, Duration		Tariff variable 3, People	
Intercept	Grouping	Factor	Grouping	Factor	Grouping	Factor
33.70506	16-30	1.000000	0-3	1.0000000	1	1.000000
	31-48	1.247261	4-10	0.9852040	2-3	1.831768
	49-59	1.277058	11-20	0.8307946	4-6	2.230428
	60-73	1.741338	21-30	0.5659430		
	74-90	1.466582				

Table 6. *Table of chosen groups and estimated GLM parameters for the truncated model data set.*

	Tariff variable 1, Age		Tariff variable 2, Duration		Tariff variable 3, People	
Intercept	Grouping	Factor	Grouping	Factor	Grouping	Factor
40.96621	16-30	1.000000	0-3	1.0000000	1	1.000000
	31-48	1.225353	4-10	0.9791169	2-3	1.698679
	49-59	1.248824	11-20	0.8474670	4-6	2.041375
	60-73	1.639699	21-30	0.5838483		
	74-90	1.384725				

Table 7. *Table of chosen groups and estimated GLM parameters for the excluded model data set.*

These values are not surprising and some trends that were expected can be seen in the table. For example, for traveling insurance it is expected that the premium does increase for people as they retire at around 60 and starts to travel more (This is represented by the value 1.6 above), but then decreases again as they get older (1.4 for the oldest age group of ages 74-90).

3.2.2.2 BY

For BY, additional plots of the building predictors were included since they may affect claim cost.

For these groups, it was decided to not include building age since the graph was very linear and didn't suggest any meaningful trend, even though it intuitively might make sense that older buildings are more prone to damages. A reasonable explanation might be that increasing frequency of damage is countered by cheaper repairs.

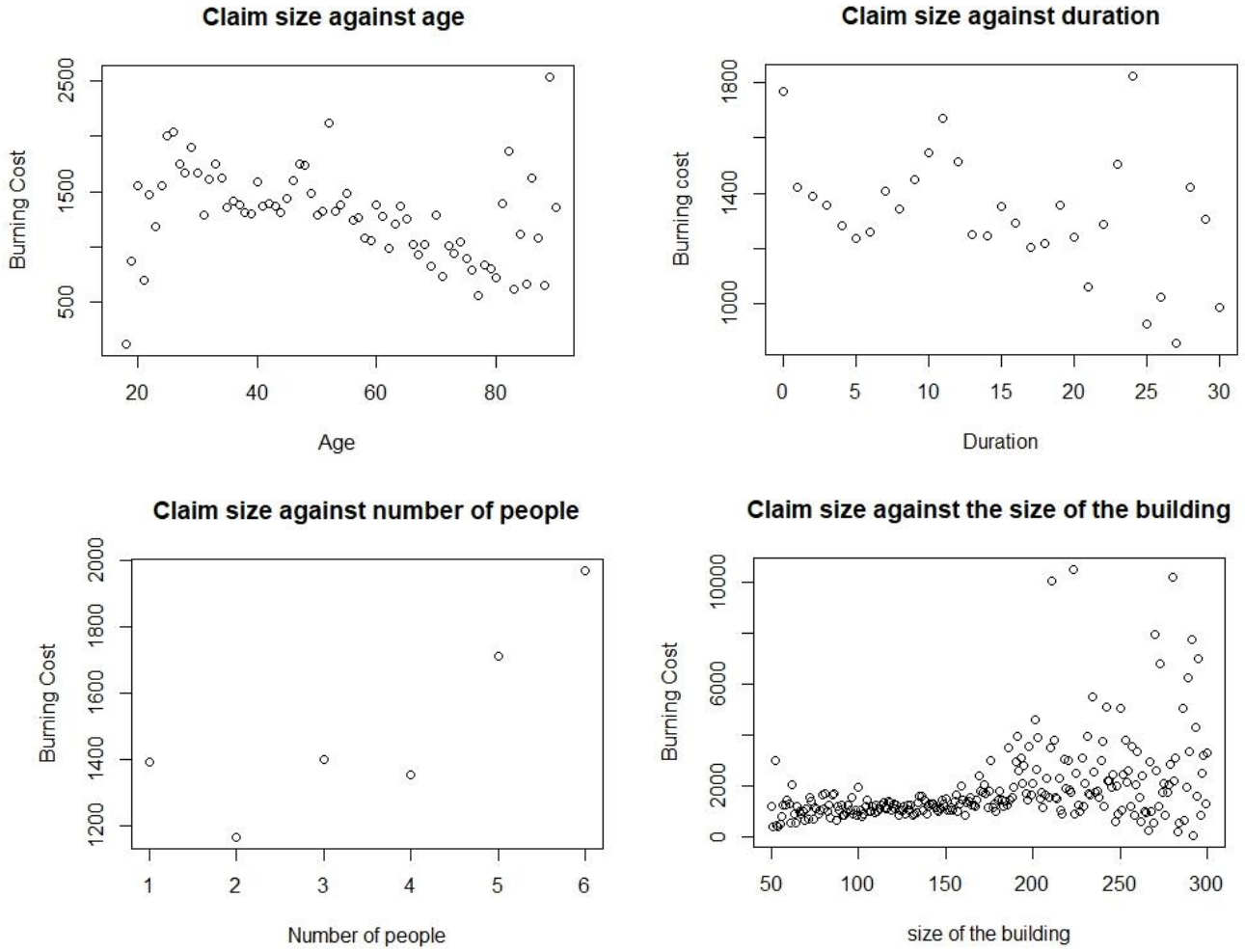

Figure 14: *Plots of burning cost against various predictor variables*

The tariff variable for building size was divided into three pieces. It appears that this parameter has a clear increase in value that occurs somewhere after $150m^2$. The exact grouping can be seen below in addition to the factors obtained from the GLM analysis with these groups as dummy variables.

Intercept
1671.44

Tariff variable 1, Age		Tariff variable 2, Duration	
Grouping	Factor	Grouping	Factor
16-30	1.000000	0-3	1.0000000
31-48	0.7991677	4-10	0.9051077
49-59	0.7094386	11-20	0.9332761
60-73	0.6156057	21-30	0.8567604
74-90	0.5656867		
Tariff variable 3, People		Tariff variable 4, Living area	
Grouping	Factor	Grouping	Factor
1-2	1.000000	50-100	1.000000
3-4	0.9709593	101-150	1.033934
5-6	1.0943083	151-300	1.343955

Table 8. *Table of chosen groups and estimated GLM parameters for the truncated model data set.*

And for exlusion:

Intercept
1499.027

Tariff variable 1, Age		Tariff variable 2, Duration	
Grouping	Factor	Grouping	Factor
16-30	1.000000	0-3	1.0000000
31-48	0.8123392	4-10	0.9157621
49-59	0.7436398	11-20	0.9277056
60-73	0.6770928	21-30	0.9051189
74-90	0.6127481		
Tariff variable 3, People		Tariff variable 4, Living area	
Grouping	Factor	Grouping	Factor
1-2	1.000000	50-100	1.000000
3-4	1.048029	101-150	1.095659
5-6	1.153999	151-300	1.339003

Table 9. *Table of chosen groups and estimated GLM parameters for the excluded model data set.*

The GLM result seems reasonable. The peak appears for the youngest age group and there is a trend of decreasing claim size with longer durations. More people on the insurance and larger buildings are both factors that increase the claim size, expectedly.

3.2.2.3 PL

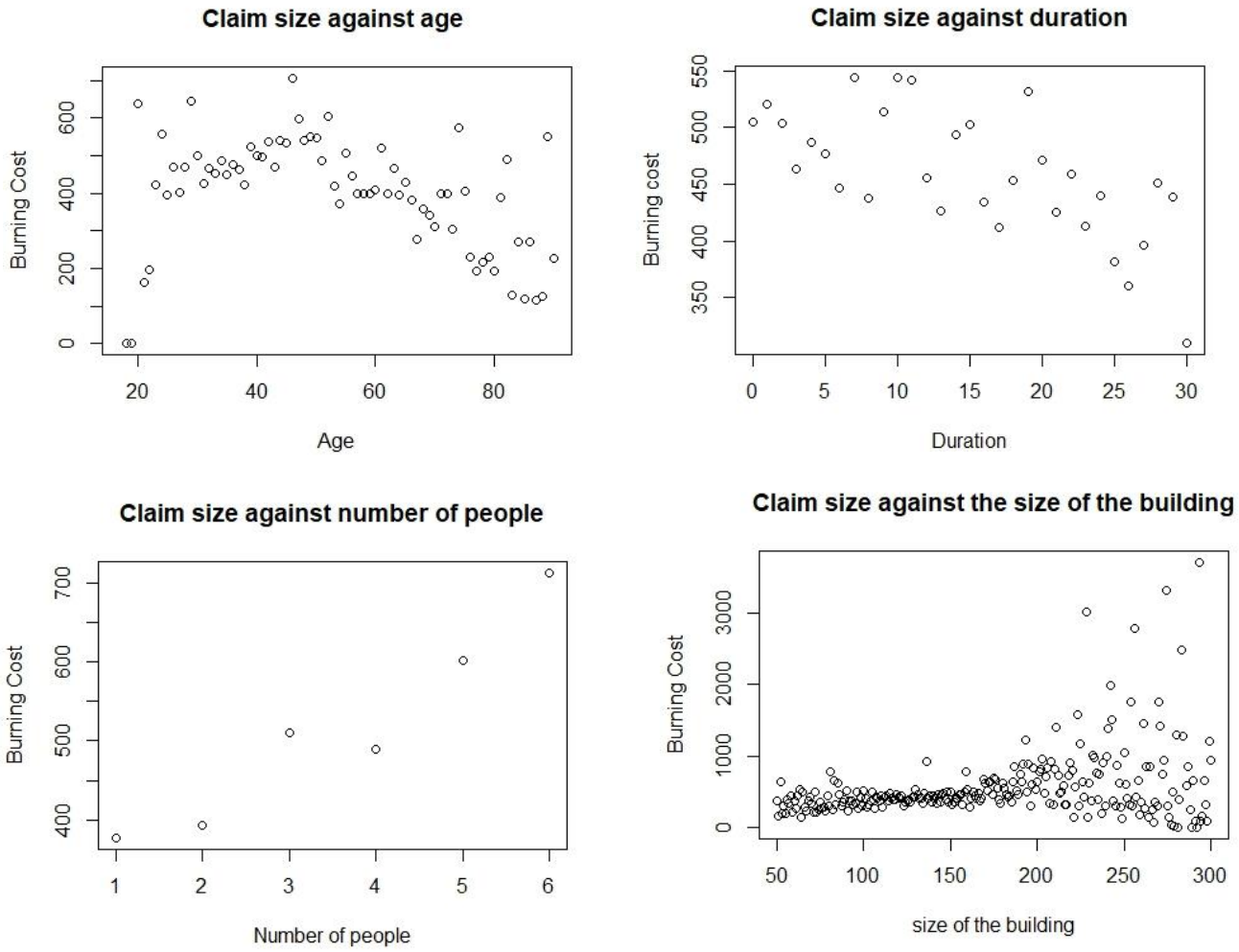

Figure 15: *Plots of burning cost against various predictor variables*

These plots were processed in the exact same fashion and lead to the following groups and GLM result:

Intercept
544.0388

Tariff variable 1, Age		Tariff variable 2, Duration	
Grouping	Factor	Grouping	Factor
16-30	1.000000	0-3	1.0000000
31-48	0.9805949	4-10	0.9644471
49-59	0.9922655	11-20	0.9454218
60-73	0.9135613	21-30	0.8740876
74-90	0.7097343		
Tariff variable 3, People		Tariff variable 4, Living area	
Grouping	Factor	Grouping	Factor
1-2	1.000000	50-100	1.000000
3-4	1.079227	101-150	1.036413
5-6	1.202845	150-300	1.228967

Table 10. *Table of chosen groups and estimated GLM parameters for the truncated model data set.*

And for exclusion:

Intercept
601.3397

Tariff variable 1, Age		Tariff variable 2, Duration	
Grouping	Factor	Grouping	Factor
16-30	1.000000	0-3	1.0000000
31-48	0.9428087	4-10	0.9787339
49-59	0.9544365	11-20	0.9579379
60-73	0.8866662	21-30	0.9053018
74-90	0.6970814		
Tariff variable 3, People		Tariff variable 4, Living area	
Grouping	Factor	Grouping	Factor
1-2	1.000000	50-100	1.000000
3-4	1.089430	101-150	1.036974
5-6	1.200729	150-300	1.168011

Table 11. *Table of chosen groups and estimated GLM parameters for the excluded model data set.*

The trends of this table are all intuitive: Increased claim size with number of people on the insurance, increased claim size with larger buildings and a descending claim size for longer durations. Finding the peak of the burning cost in the youngest age group is also not surprising.

3.2.3 Champion model

The result of the champion model is presented by the burning cost of the model against the true value of the burning cost from the reference data. The model that describes the true burning cost with the most accuracy is chosen. In each product code, one comparisons between

truncation and exclusion are made, the threshold limit used is the one obtained from the previous results.

The certainty of a given point in the graph below is given by the dot that represents the exposure. The points with very little exposure must be taken with a grain of salt since the error can be very large in those points. So, when determining which method that fits the best, more weight is put into the points with high exposure.

Below are the graphs for each product code.

3.2.3.1 RP

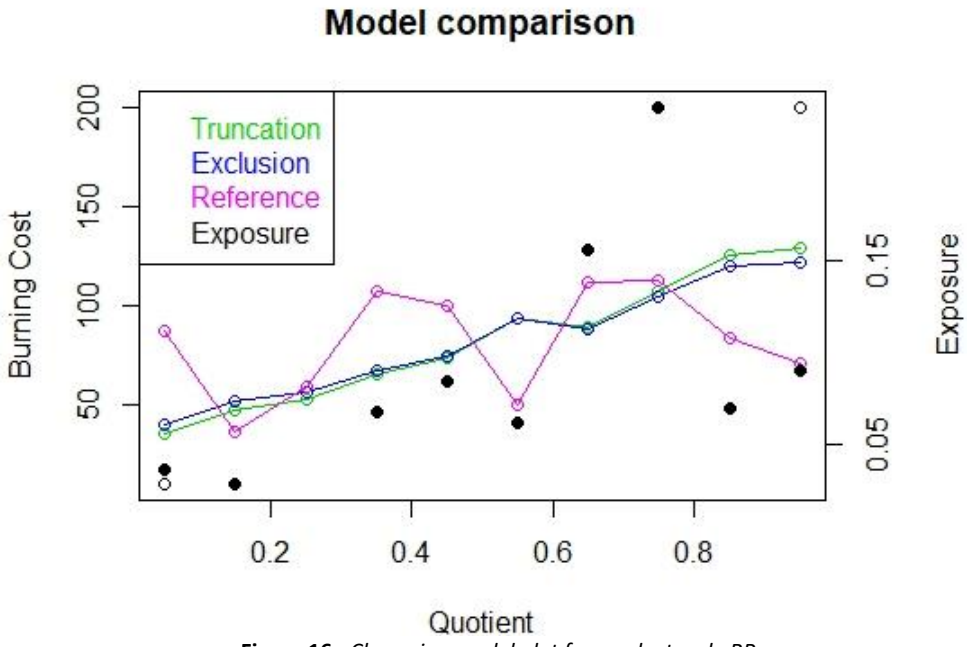

Figure 16: *Champion model plot for product code RP*

For the 24 000-threshold limit the Exclusion model fits the reference data with the most accuracy. Note that Quotient Ranking is not equivalent to the value of the quotient. Each point on the horizontal axis, say 0.1, represents 10% of the quotients (the lowest ones at 0.1, lumped together).

3.2.3.2 BY

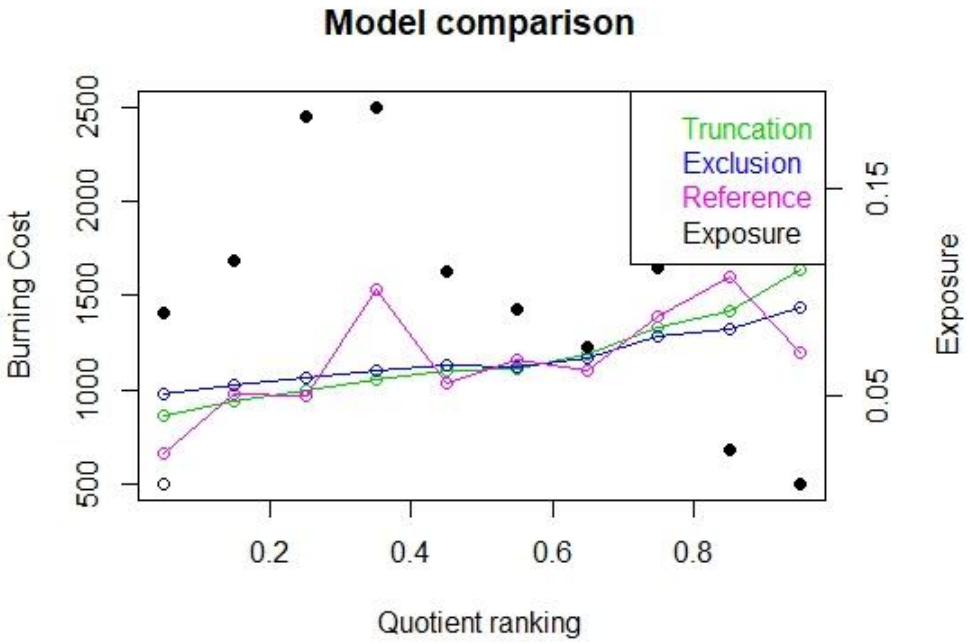

Figure 17: *Champion model plot for product code BY*

The 800 000 plot suggests truncation being the better choice of how to treat the large claim limit. The model do have similar performance so either could probably be chosen without any trouble – but still the truncation model comes out with a slight edge.

3.2.3.3 PL

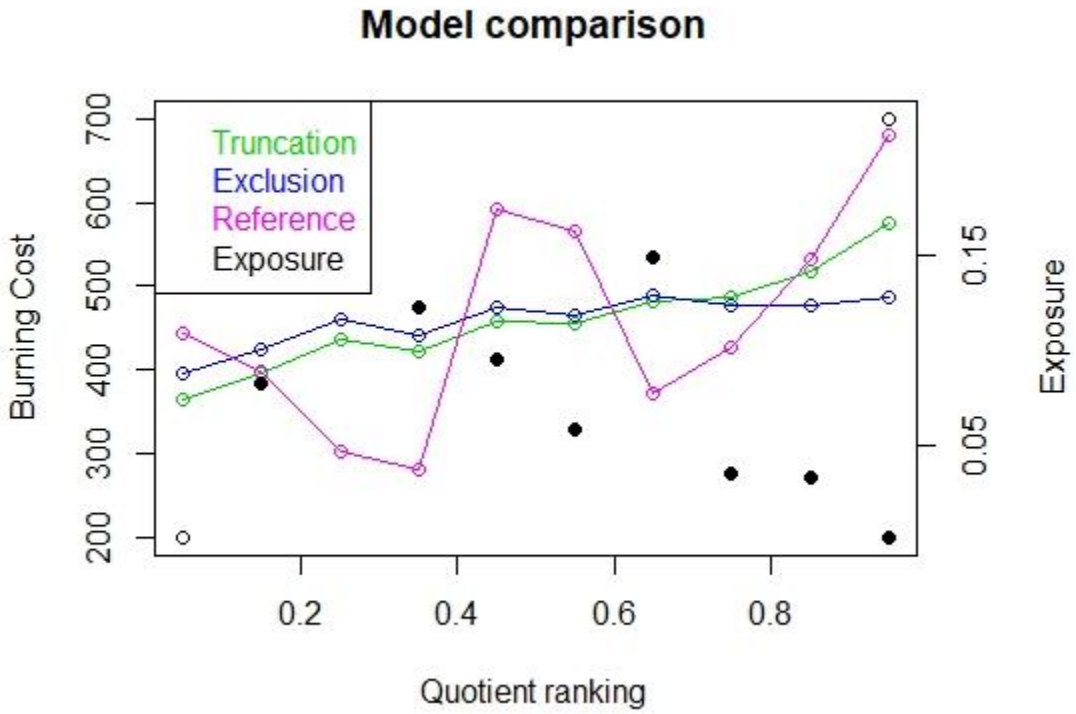

Figure 18: *Champion model plot for product code PL*

For the 275 000 limit, truncation once again seems to come out on top. In the majority of the points where the models do not accurately describe the reference data, truncation is closer most of the time. In the other cases they appear to have similar performance.

4 Conclusion and Discussion

The result with the modified data set for the problem can be summarized by the following table.

Product code	Large claim limit	Truncation/exclusion
RP	24 000	Exclusion
BY	800 000	Truncation
PL	275 000	Truncation

Table 12. *Summary of the result*

As seen in the table the large claim limit varies very much between the product codes. These differences do however make sense intuitively. You would expect the claim size of a building damage to be much greater than that of a personal property loss/damage. And with a much greater mean claim size, the limit that denotes a large claim will naturally increase with it.

4.1 Mean residual life & Parameter stability

In each individual plot it is difficult to find an exact interpretation for the threshold limit. However, by finding the limit that suits parameter stability and mean residual life plots at the same time, that interpretation is made easier. But keep in mind that you most certain could use a large claim limit that is a bit higher or a bit lower than the presented large claim limit and still get the same result.

4.2 Champion model

As in any regression, while making the groups for the GLM analysis the goal was to create as many groups as possible to not miss any relevant behavior of the predicting variables.

A slight problem for the Champion model can also occur with small groups. GLM handles small groups quite well when predicting values. However, when comparing the estimated values of the GLM with the true value of small groups, there is very little to no data for some of the groups. For example, there is no exposure at all for the youngest age group with 20 years of insurance duration (naturally, since this combination should be impossible). This gives the reference data very inexact or missing values at some of the quotients in the result plots. However, there are fortunately plenty of groups with high exposure that the model performance can be compared more easily.

Some prediction variables were not included at all in the champion model for certain product codes. For example, building age was not a part of the regression in RP. One could argue that information about the building could be relevant even in the case of travel insurance, but that was a nuance that was estimated to be small enough to be overlooked.

4.3 Choosing Threshold

The methods used to determine the threshold limit were mainly graphical. Naturally with graphical methods come subjectivity and inaccuracy. To counteract this one can increase the number of tests. This did take shape in the form of the rules of thumb. However, these methods

did have less theory behind them and consequently the qq and pp plots suggested the limits from the graphical methods as a better fit.

4.4 Weighing Truncation against Exclusion

The results regarding truncation vs exclusion on the exceedances over the threshold limit, does depend on product code. This is not surprising since the codes separate different kinds of damage which may have different underlying distributions.

For BY and PL, truncation seemed to be the better fit, while for RP exclusion seemed to be better, however not by a longshot. In fact, in neither case, either method is clearly better or worse than the other.

However, even though the favor of truncation given by the results in this thesis is only minor – intuitively it may also be a better option. A good example of why is that for exclusion, data points just above a threshold limit, let's say 801 000, would be totally excluded from the GLM analysis, while claims just below, say 799 000, would be included. Even though these two data points has practically the same value, one makes it in and the other doesn't.

For a truncation type model, this problem obviously doesn't occur.

4.4 Future work

In many areas of this thesis we felt that the analysis could be improved if there was more data. Some of the final groups in the GLM analysis still has very little to no exposure so comparing the exclusion and truncation models with the true value ended up being nearly impossible due to some of the groups not having reliable or any data to compare to from the reference data set. One could argue though, that it is not important to be able to compare the model's performance in those groups, for example 80-year-old people living 5 in the same house. In addition, truncation and exclusion did perform very similarly to each other. The true value might not be relevant for the scope of this thesis if one can already conclude that the method of choice above the threshold limit gives roughly the same results.

In this thesis, the cost that exceeds the threshold limit is spread evenly over all of the insurance holders. This might be unfair since the cheapest and the most expensive insurance holder groups have the same value added to their insurance costs. This could be a point worth investigating. One could compare the performance of the current praxis with a model that puts a larger part of the exceeding claims on the costs of the more expensive groups.

5 References

Ananda, M. Cooray, K. (2007). Modeling actuarial data with a composite lognormal-Pareto model

Coles, S. (2001). An Introduction to Statistical Modeling of Extreme Values, springer. Pages 74-

Davison, A.C. & Smith, R.L. (1990). Models for Exceedances over High Thresholds. Journal of the Royal Statistical Society. Series B (Methodological), Vol. 52, No. 3. Pages 393-442.

Embrechts, P. Klüppelberg, C. Mikosch, T. (2012). Modelling Extremal Events for Insurance and Finance, Springer. Pages 152-168,352-356.

Ferreira, A. de Haan, L. & Peng, L. (2003). On optimising the estimation of high quantiles of a probability distribution, Statistics, 37. Pages 401–434.

Hult, H. Lindskog, F. et al. (2012). Risk and Portfolio Analysis Principles and Methods, Springer. Pages 265-270.

Johansson,B & Ohlsson,E. (2010). Non-Life Insurance Pricing with Generalized Linear Models, Springer. Pages 1-35

Loretan, M. & Philips, P.C.B. (1994). Testing the covariance stationarity of heavy tailed time series: an overview of the theory with applications to several financial datasets, J. R. Statist. Soc. D, 1. Pages 211–248.

Mcdonald, A. Scarrot, C. (2012). A review of extreme value threshold estimation and uncertainty quantification. Volume 10, number 1. Pages 33-60