Erasmus University Rotterdam

Bachelor Thesis in Financial Econometrics

Extremal Random Forest Model to Predict Extreme Quantiles in Medical Claim Data

Author

Gijs Jacobs (506 679)

Supervisor

Prof. dr. Chen Zhou

Second Assessor

Aishameriane Venes Schmidt

July 3, 2022

Abstract

For insurers, extremes in medical claim data are important to predict. In this paper, we try to predict extreme quantiles of large medical claim data and try to investigate their determinants. Gnecco et al. [2022] provides the extremal random forest (ERF) method to predict extreme quantiles by means of random forests. We evaluate the performance of this ERF method compared to the generalized random forests (GRF) method proposed by Athey et al. [2019] and the unconditional generalized Pareto distribution (GPD) method (Tajvidi [2003]). To perform a model comparison by means of predictor performance and robustness, we perform a simulation study. Furthermore, the methods are applied to a medical large claim data set containing over 95,000 health claims of $25,000 and above and an U.S. wage data set. We find that the ERF method outperforms the other methods for the U.S. wage data set. However, for the medical large claim data, ERF does not outperform the GRF and unconditional GPD method.

The views stated in this thesis are those of the author and not necessarily those of the superviser, second assessor, Erasmus School of Economics or Erasmus University Rotterdam.

Contents

List of Figures

List of Tables

1 Introduction

For health insurers, it is of importance to estimate the likelihood of medical claims. The number of claims and the height of the reimbursement of these claims determines their total reimbursement expenditures (Casto and Forrestal [2013]). Especially, large medical claims are of great importance due to their impact on the buffer for the insurer. Insurers can use predictive information on large claims to ensure they have enough in cash and do not have to make unwanted changes in their portfolio.

Therefore, we make use of quantiles. Quantiles are cut points dividing observations in a sample into continuous intervals with equal probabilities. According to Yaqiong [2018], the focus of risk management in insurance companies should be in health insurance. This is due to the relative large impact extreme claims in health insurance have compared to extreme claims in life and accident insurance. Therefore, our goal is to predict extreme quantiles of medical claims and investigate their determinants.

To make predictions, external information can be used in a conditional quantile regression. This is a linear model in which the quantile of the response variable is estimated based on external information. Chernozhukov [2005] develops a theory of quantile regression in the tails. Specifically, the large sample properties of extremal (extreme order and intermediate order) quantile regression estimators for the linear quantile regression model are obtained.

All possible explanatory variables are in the predictor space. When there are multiple explanatory variables in the predictor space, a high dimension is present. To cope with the high dimension of the predictor space, several methods have been proposed based on random forests originally developed by Breiman [2001]. Meinshausen [2006] showed that random forests provide information about the full conditional distribution of the response variable, not only about the conditional mean. Later, Athey et al. [2019] proposed an algorithm to grow generalized random forests (GRF), which were able to make quantile prediction for more complex quantile surfaces in larger dimensions. These methods work well for estimation of quantiles inside the data range, so for intermediate quantile levels. However, these methods do not work well for extreme quantile levels in the right tail.

In the field of economics, Gnecco et al. [2022] performed an empirical study on extreme quantile prediction for U.S. wage data. They introduced a new extremal random forest (ERF) algorithm that combines both the advantages of accurate tail extrapolation at extreme levels with a flexible regression method that scales well with predictor dimension. To obtain accurate tail extrapolation, we make use of the generalized Pareto distribution (GPD). The GPD is an often used distribution to model exceedances over a high threshold (Tajvidi [2003]). It is specified by two parameters: a scale and shape parameter. This ERF algorithm makes use of forest-based weights to localize the estimation of the the scale and weight parameters.

In summary, we investigate performance of the ERF method when applied to medical claim data. Our main focus is in the determinants of large claims. We evaluate the performance of the ERF method on large claims data. Thus, our research question is: *'How is the performance of the ERF method compared to other quantile regression methods for medical large claim data?'* Furthermore, the ERF method will be applied to 1980 U.S. wage data and its performance will be compared against other methods. We find that the ERF method outperforms both the classical GRF method as the unconditional GPD method for the U.S. wage data. While, for the medical claim data ERF does not outperform the GRF and unconditional GPD method. So, the combination of accurate tail extrapolation with a flexible regression method that scales well with predictor dimension performs best against more 'classical' methods for the U.S. wage data but not for the medical claim data.

The paper proceeds as follows. Section 2 covers the literature on this topic. Section 3 provides the data diagnostics. In Section 4, we elaborate on the methods used. The simulation study will be discussed in Section 5 followed by the application results of U.S. wage data and medical claim data in section 6. Finally, our results and study will be concluded and discussed in Section 7.

2 Literature Review

Extremes in insurance claims cause fat tails, therefore classic density functions do not fit all the data. Therefore, Uwingabire [2018] modelled extreme health insurance claims using generalized Pareto distribution. Another approach to deal with these extreme observations is to solely focus on these extreme observations (Balasooriya and Low [2008]). Consequently, they used a highly flexible distribution to cope for extremes, the generalized lambda distribution. Another model used by Balasooriya and Low [2008] is the semiparametric transformed kernel density estimation. They found that both approaches performed well compared to the generalized Pareto distribution. Cebrián et al. [2003] proposed a statistical modeling strategy based on extreme value theory with main focus on large claims, as in our research. Their proposed strategy is compared to several standard distributions as the gamma, lognormal, and log-gamma distribution.

Gnecco et al. [2022] propose the ERF method which relies on approximation by the GPD to extrapolate exceedances over a certain threshold. They show by means of simulations that their ERF method performs better than classical quantile methods and existing regression approaches based on extreme value theory. Our paper gives new insights in the application of the ERF method proposed by Gnecco et al. [2022]. They showed that the ERF method is a compatible method for extreme quantile wage estimation with high dimension predictor space. We investigate if this ERF method works well for lower dimension space when applied to medical claim data. Our focus is mainly on health economics and will be relevant to insurance companies and governments.

3 Data

We perform two different applications, both with a different data set. First, we compare the performance of ERF, GRF and unconditional GPD on 1980 U.S. wage data from Angrist et al. [2009]. This data consist of 65,023 black and white men in the age 40-49, with an education level between 5 and 20 years. The numeric covariates 'Age', 'Experience' and 'Education' and the binary covariate 'Black' to indicate whether the person is black are used. Summary statistics of the weekly wage, calculated as the annual wage divided by the number of weeks worked in the previous year are presented in Table 1. For our health claim analysis we will use the 1992 Medical Insurance Large Claims Database from the Society of Actuaries (SOA).[1] This database includes over 95,000 claims of $25,000 and above, representing over $5 billion in total charges. The covariates include 'Gender', representing whether someone is a male or female and the numeric covariate 'Age'. Some descriptive statistics for the total of the claims in $ are represented in Table 2. This data consist of 49,242 men (51.6%) and 46,189 women (48.4%).

Mean	600.31
Minimum	0.16
Maximum	80,161.61
Median	634.17
1st Quantile	469.40
3rd Quantile	852.58

Table 1: Summary statistics for weekly wages (in $).

Mean	58,791
Minimum	25,000
Maximum	7,104,081
Median	40,105
1st Quantile	30,498
3rd Quantile	61,014

Table 2: Summary statistics for claims (in $).

4 Methodology

In our research, we are interested in estimating high conditional quantiles of a response variable $Y \in \mathbb{R}$ across values of the predictor variables $X \in \mathbb{R}^p$ with a high dimension p. In order to model the distribution of the tails, the following GPD function from Pickands [1975] is used,

$$G(z;\theta(x)) = 1 - \left(1 + \frac{\xi(x)}{\sigma(x)} z\right)_+^{-1/\xi(x)}, \quad z > 0, \tag{1}$$

where $\sigma(x)$ is the scale parameter and $\xi(x)$ is the shape parameter. Combining Equation (1) and Bayes' theorem, Gnecco et al. [2022] obtained the GPD approximation for small probability of Y exceeding a high threshold y. This approximation is given in Equation (2).

$$\mathbb{P}(Y > y) = \mathbb{P}(Y > u)\mathbb{P}(Y > y \mid Y > u) \approx \mathbb{P}(Y > u)\{1 - G(y - u; \sigma(x), \xi(x))\}, \tag{2}$$

where u is an intermediate threshold, for which holds $u < y$.

Gnecco et al. [2022] denotes $Q_x(\tau)$ as the conditional quantile at level $\tau \in (0, 1)$ of Y, given $X = x$.

[1] https://www.soa.org/resources/experience-studies/2000-2004/91-92-group-medical-claims/

As our interest is in estimating extreme quantile levels, we focus on τ close to one. Following from Balkema and de Haan [1974] and Pickands [1975] the conditional quantile of Y, given $X = x$ for extremal quantile levels can be approximated by

$$Q_x(\tau) \approx Q_x(\tau_0) + \frac{\sigma(x)}{\xi(x)} \left[\left(\frac{1-\tau}{1-\tau_0} \right)^{-\xi(x)} - 1 \right], \tag{3}$$

where $Q_x(\tau_0)$ is an intermediate quantile at τ_0, for which hold $\tau_0 < \tau$. Equation (3) is obtained by combining Equation (1) and (2).

Equation (3) consists of two parts to be estimated, the intermediate level quantile conditional on $X = x$ given by $Q_x(\tau_0)$. The second part of interest on the right hand side determines extrapolated data approximated by the generalized Pareto distribution (GPD). This requires estimations for the conditional scale $\sigma(x) > 0$ and shape parameter $\xi(x) \in \mathbb{R}$.

Following the ERF algorithm from Gnecco et al. [2022], we first estimate $Q_x(\tau_0)$, the similarity weights function $w_n(x, y)$ is estimated by GRF from Athey et al. [2019]. Secondly, the intermediate quantile function $Q_x(\tau_0)$ is estimated, for which hold $\tau > \tau_0$ where τ_0 and τ are the intermediate and extreme quantile levels, respectively. $\hat{Q}_x(\tau_0)$ is fitted by a classical quantile regression technique.

For the second part of the ERF algorithm, one uses the exceedances to predict the extreme quantile $\hat{Q}_x(\tau)$. As in Gnecco et al. [2022], we minimize the weighted (negative) log-likelihood in Equation (4) to obtain the estimator of the conditional GPD parameter $\theta(x) = (\sigma(x), \xi(x))$.

$$L_n(\theta; x) = \sum_{i=1}^{n} w_n(x, X_i) \, \ell_\theta(Z_i) \, \mathbb{1}\{Z_i > 0\}, \quad x \in \chi. \tag{4}$$

In this equation Z_i are the exceedances in the training data defined as $Z_i := (Y_i - \hat{Q}_{X_i}(\tau_0))_+, i = 1, \ldots, n$. As in Gnecco et al. [2022] we use GRF with quantile loss to fit $\hat{Q}_x(\tau_0)$ due to its convenience for high-dimensional quantile regression problems and only little tuning is needed. Moreover, $\ell_\theta(Z_i)$ is the negative log-likelihood contribution of the ith exceedance Z_i which is defined in Equation (5) if $Z_i > 0$.

$$\ell_\theta(Z_i) = \log \sigma + \left(1 + \frac{1}{\xi}\right) \log \left(1 + \frac{\xi}{\sigma} Z_i\right), \quad \theta \in (0, \infty) \times \mathbb{R}, \tag{5}$$

if $Z_i < 0$, $\ell_\theta(Z_i)$ is equal to zero.

Subsequently, the obtained parameters and variables are substituted in Equation (3) resulting in the estimated extreme quantile $\hat{Q}_x(\tau)$. As can be conducted from the algorithm in Gnecco et al. [2022], only in the fitting part, random forest is used.

Tuning parameters such as the minimum node size and the number of predictors selected at each split are of great importance within the ERF algorithm. We tune these hyperparameters by the following cross-validation scheme provided by Gnecco et al. [2022]. First, we partition $\{1, \ldots, n\}$ into M equally sized folds of the training data. Thus $N_1, \ldots, N_M$ will be the random partitioning. Hereafter, we estimate the GPD parameter vector $\hat{\theta}(X_i; \alpha_j)$. Where $\alpha_j \in \{\alpha_1, \ldots, \alpha_J\}$ is a sequence of tuning parameters that we fit on the training set (X_i, Y_i). Finally, we calculate the cross-validation error by

$$CV(\alpha_j) = \sum_{m=1}^{M} \sum_{i \in N_m} \ell_{\hat{\theta}(X_i;\alpha_j)}(Z_i) \, 1\{Z_i > 0\}. \tag{6}$$

The value of α for which $CV(\alpha_j)$ is minimized, will be the optimal tuning parameter. Following Gnecco et al. [2022] the minimum node size κ plays the most critical role for ERF. This is due to its property to control the model complexity of the individual trees in the forest. As a consequence, κ controls the complexity of the similarity weights $w_n(x, y)$ too.

As a result of Equation (3), tail behaviour of Y at extreme quantile levels is determined by the shape ξ of the GPD. Since estimation of the shape parameter is difficult, following Hastie et al. [2009] showed that using penalization it is possible to lower an estimator's variance at the expense of increased bias. Several penalization schemes are proposed (Smith and Naylor [1987], de Zea Bermudez and Turkman [2003]) based on i.i.d. data. While our interest is in penalizing the shape function $\xi(x)$ across the predictor space. Gnecco et al. [2022] proposed to penalize the weighted GPD (4) with the squared difference between the estimates of $\xi(x)$ and the constant shape parameter ξ_0, resulting in

$$\hat{\theta}(x) = \underset{(\sigma,\xi)=\theta\in\Theta}{\arg\min} \frac{1}{(1-\tau_0)} L_n(\theta; x) + \lambda \left(\xi - \xi_0\right)^2, \tag{7}$$

where ξ_0 is obtained by minimizing Equation (4) for equal weights and λ is a tuning parameter. In the next section we perform a simulation study including cross-validation and penalization results.

5 Simulation Study

We run a simulation study to compare the performance of ERF to other quantile regression methods. Competing methods of interest will be the generalized random forest (GRF) by Athey et al. [2019] and as a baseline, the unconditional GPD model is considered.

5.1 Simulation Setup

We use the same setup as in Gnecco et al. [2022]. For the simulation study we simulate n independent copies of a random vector (X, Y), $(X_1, Y_1), \ldots, (X_n, Y_n)$ which are the training observations. Let $X \in \mathbb{R}^p$ be a uniform distribution on the cube $[-1, 1]^p$ for dimension p. The response variable $Y|X = x$ follows a Student's t-distribution with 4 degrees of freedom and scale $s(x) = 1 + \mathbb{1}\{x_1 > 0\}$. The goal of this simulation study is to investigate the effect of different quantile levels (for a fixed dimension) on the prediction performance for the competing methods. Furthermore, we investigate differences in prediction performance between competing methods for increasing dimensions (for a fixed extreme quantile level).

The second experiment is to investigate robustness of the quantile regression methods. For a large, given dimension p, we simulate data with shape parameters $\xi = 0, 1/4, 1/3$. For the case $\xi = 0$, a Gaussian distribution is chosen. For $\xi = 1/4, 1/3$ a Student's t-distribution is chosen with respectively 4 and 3 degrees of freedom.

5.2 Performance Measure

As in Gnecco et al. [2022] we evaluate the performance of the method on a test data set, generated with a Halton sequence (Halton [1964]) on the cube $[0,1]^p$. The test data set consists of $m = 1000$ observations. We then compute the integrated sequence error (ISE) on the test data following Equation (8).

$$\text{ISE} = \frac{1}{m}\sum_{i=1}^{m}\left(\hat{Q}_{x_i}(\tau) - Q_{x_i}(\tau)\right)^2, \tag{8}$$

where $\hat{Q}_{x_i}(\tau)$ is the fitted quantile regression function, and $Q_{x_i}(\tau)$ is the true quantile function. Repeating the simulation, fitting and calculation of (8) 50 times and averaging results in the mean integrated squared error (MISE). Performance of the three methods will be measured by the MISE.

5.3 Hyperparameter Tuning

In Figure 1, the square root of MISE is showed for different minimum node sizes κ on three different quantile levels τ. We set $n = 2000, p = 40$ and $\tau_0 = 0.8$. We make use of 5-fold cross-validation.

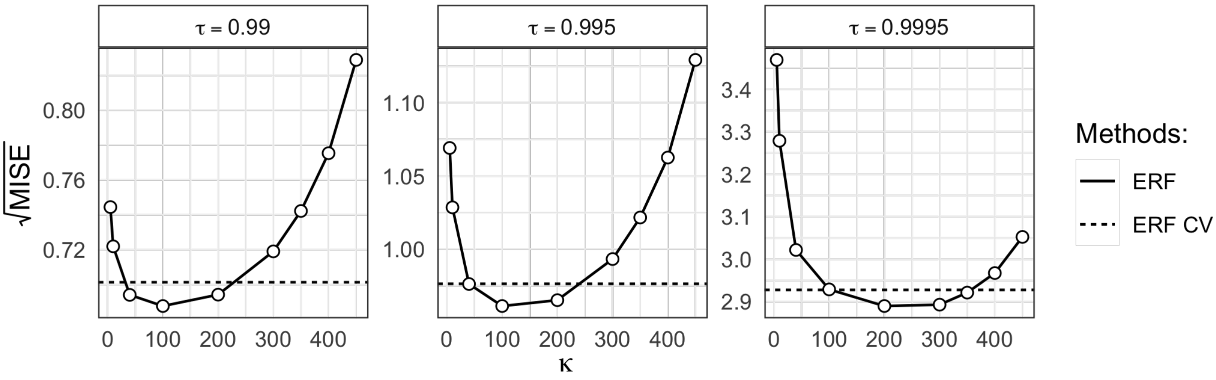

Figure 1: Square root MISE of ERF for various minimum node sizes κ and quantile levels τ over 50 simulations. The square root MISE of the cross-validated ERF is represented by the dashed line.

As can be observed in Figure 1 the dashed line is close to the minimum square root MISE, meaning the proposed cross-validation of ERF works well.

5.4 Penalized Log-Likelihood

In Figure 2, square root of MISE of ERF is shown over 50 simulations for different quantile levels $\tau = 0.99, 0.995, 0.9995$ and different levels of $\lambda \in \{0, \ldots, 0.1\}$.

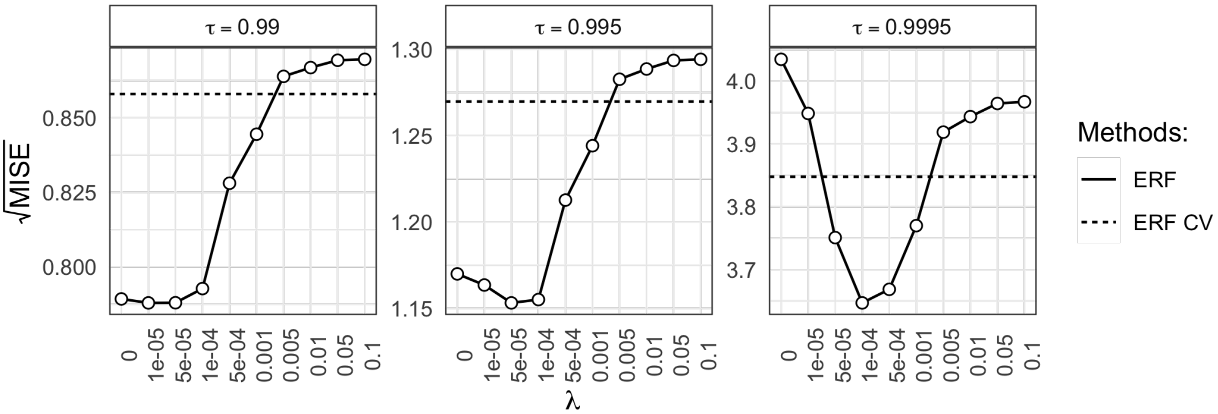

Figure 2: Solid line represents square root of MISE of ERF for different penalty values λ and quantile levels τ over 50 simulations. The dashed line represents square root MISE of cross-validated ERF.

We observe that the dashed line is not near the minimum square root MISE. Though, for higher values of τ the dashed line gets closer to the minimum suggesting the proposed cross-validation scheme works works well.

5.5 Model Comparison by Predictor Performance

This simulation study consists of two parts. First, we are interested in the effects of different quantile levels (for a fixed dimension) on the prediction performance for the competing methods. The second part of this simulation study focuses on differences in prediction performance for increasing dimensions between competing methods. In this part, the quantile level is fixed. The competing methods in these studies are the ERF, GRF and unconditional GPD. We use the same setup as in Gnecco et al. [2022]. In the left panel of Figure 3 we fix the dimension $p = 10$ and compare the prediction performance between the different methods for different quantile levels τ. For intermediate quantile level $\tau = 0.8$, similar prediction performances are measured between the different methods. Especially, all three methods have the same performance for the intermediate quantile because they use the same GRF based estimator when $\tau = 0.8$. Though, as quantile level τ increases, the square root MISE of the three methods differ from each other. The ERF method outperforms both GRF and Unconditional GPD, especially for τ close to 1.

In the right panel of Figure 3, for increasing dimensions p of the predictor space, square root MISE can be found. The quantile level is fixed $\tau = 0.9995$. Unconditional GPD and ERF look relatively robust against growing dimensions. For the Unconditional GPD, this can be explained by the fact that this model does not use the predictor values. While, GRF has some spikes and therefore seems less robust against additional noise variables. ERF performs well for large dimensional predictor spaces and is resistant to additional noise variables, making it a robust model.

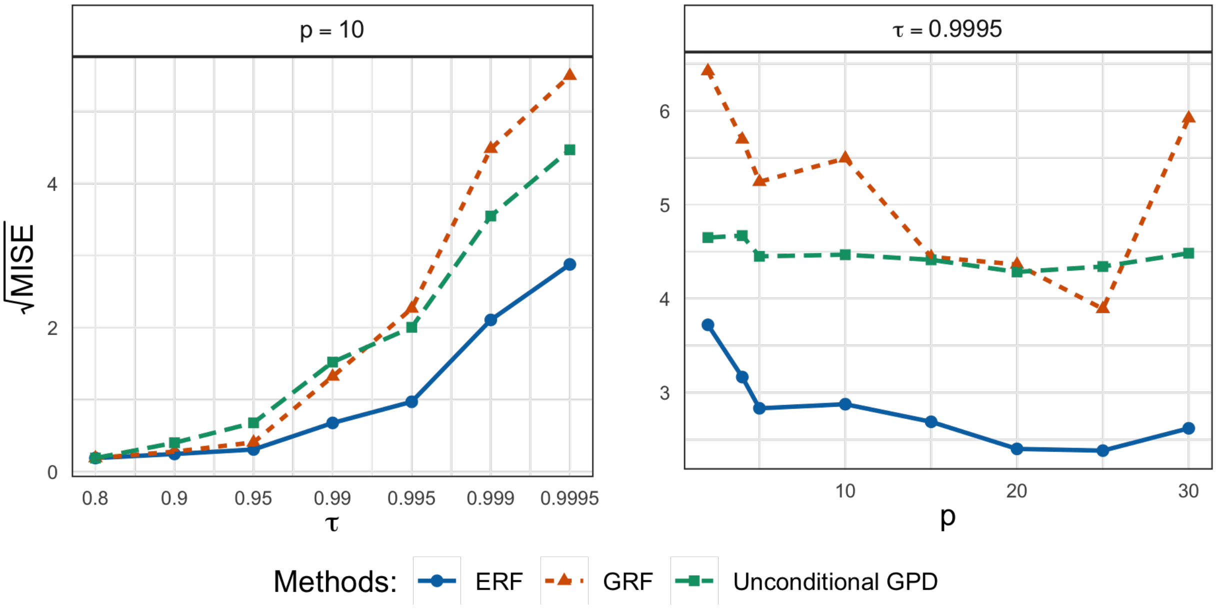

Figure 3: Left panel: Square root of MISE of three models against the quantile level τ for $p = 10$. Right panel: Square root of MISE of three models against the model dimension p for $\tau = 0.9995$.

5.6 Model Comparison by Robustness

This experiment is to investigate the robustness of quantile regression methods against noise distributions with varying tail heaviness. For a large, given dimension p, we simulate data with shape parameters $\xi = 0, 1/4, 1/3$. For the case $\xi = 0$, a Gaussian distribution is chosen. For $\xi = 1/4, 1/3$ a Student's t-distribution is chosen with respectively 4 and 3 degrees of freedom. Boxplots of the square root ISE for different shape parameters $\xi \in \{0, 0.25, 0.33\}$ for $\tau = 0.9995$ are shown in Figure 4. The means are illustrated by triangles. We observe that ERF outperforms unconditional GPD and GRF, even in the Gaussian ($\xi = 0$) case.

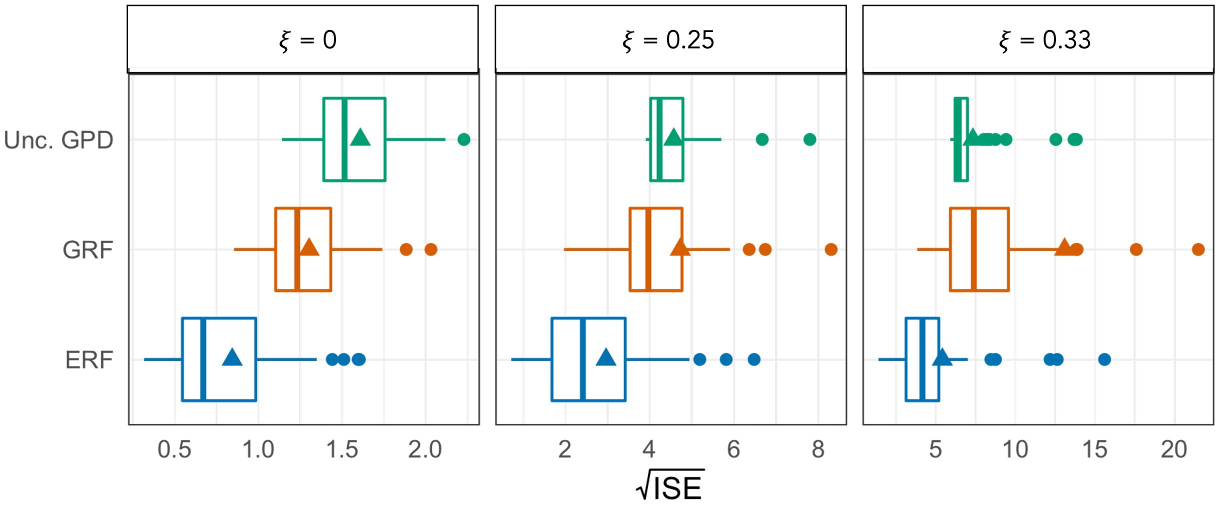

Figure 4: For 50 simulations, boxplots of the square root ISE for different values of ξ at $\tau = 0.9995$. Mean values are represented by triangles.

6 Application

We apply the ERF method to two different data sets. First, we investigate U.S. wage data whereafter we will apply the ERF method to medical claim data to compare the performance of ERF against GRF and the unconditional GPD method.

6.1 Analysis of U.S. Wage Structure

For the 1980 U.S. wage data, the response variable Y is the weekly wage, and the predictor vector consists of age and years of education and the binary variable whether a person is black or not. There are added 10 dimensions which are uniformly distributed on the interval $[-1, 1]$. This makes the total predictor space dimension $p = 13$. As in Gnecco et al. [2022] we perform a 5-fold cross-validation, three times to tune the minimum node size $\kappa \in \{5, 40, 100\}$ and we set the penalty $\lambda = 0.01$. The intermediate quantile is predicted using GRF with $\tau_0 = 0.8$ and data is split in half where the first sample is used for a data exploratory analysis and the second half will be used for fitting and evaluation of the methods. From the first sample, a 10% random subset is made which will be fitted by ERF. This predicts $\hat{\tau}(x) = (\hat{\sigma}(x), \hat{\xi}(x))$ on the remaining 90% observations of the first sample.

The estimated GPD parameters as a function of the years of education are shown in Figure 5. It can be observed that the scale parameter $\hat{\sigma}(x)$ is positively correlated to the years of education, while the shape parameter $\hat{\xi}(x)$ is negatively correlated to the years of education. Moreover, we see a shift around 15 to 16 years of education, this can be explained by the completion of a Bachelor's study. For both parameters, it is relatively homogeneous between the black and white group. For the scale parameter $\hat{\xi}(x)$ we see a stable correlation against the years of education, whereas for the scale parameter $\hat{\sigma}(x)$ some kind of exponential growth can be observed between 5 and 15 years of education.

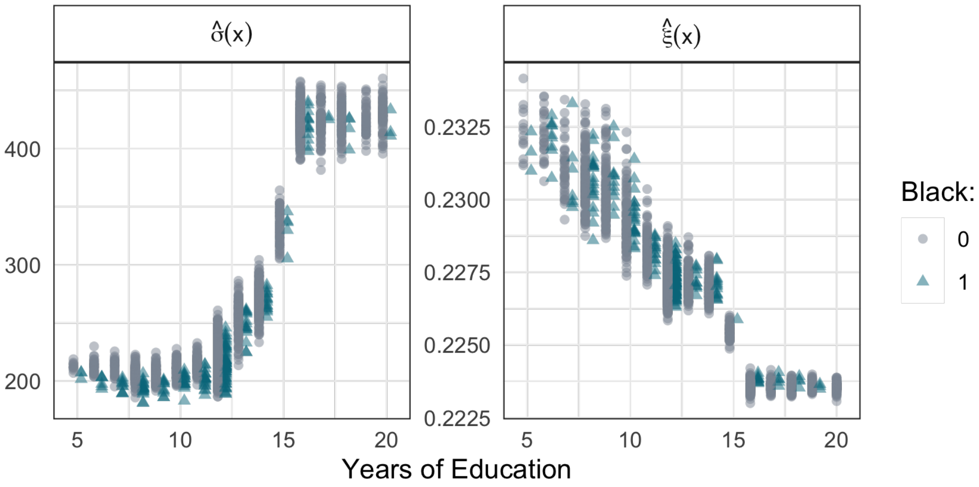

Figure 5: Estimated GPD parameters as a function of years of education.

The methods are compared against each other in Figure 6 for quantile levels $\tau = 0.9$ and $\tau = 0.995$. It can be seen that the shape of ERF and unconditional GPD method is about equal for the different levels of τ, though the unconditional GPD has a flatter pattern for higher τ. This is due to the unconditional GPD property that it cannot produce different scale parameters, whereas Figure 5 confirms that it is mandatory for this data set. When looking at GRF, the shape for τ approximately equal to 1 is very different from $\tau = 0.9$ confirming GRF is a good method for intermediate quantile approximation but not for extrapolation.

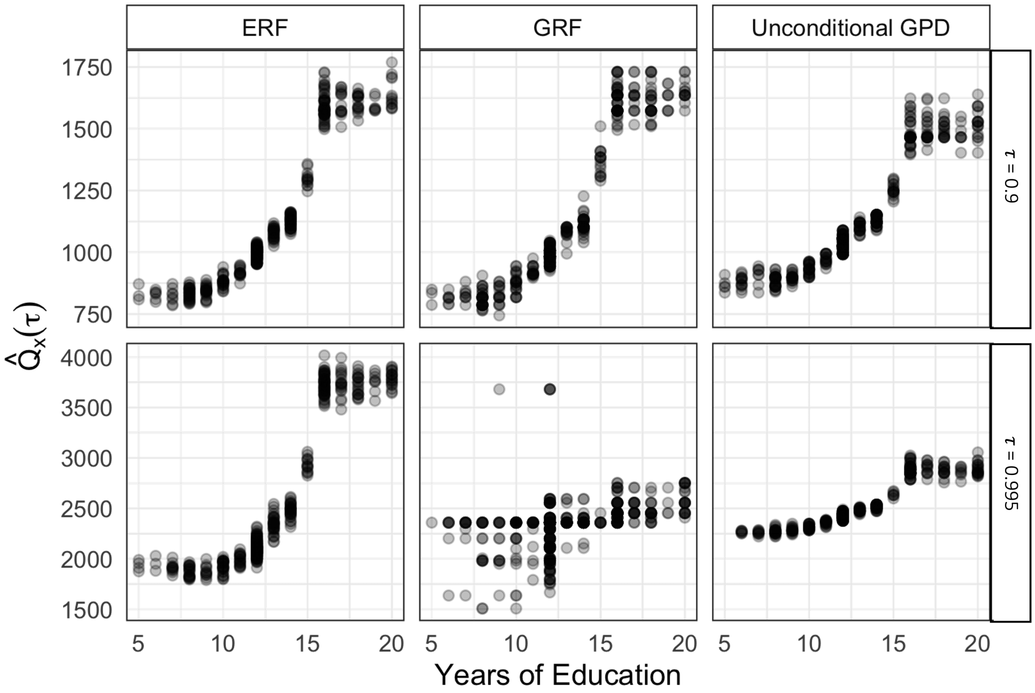

Figure 6: Predicted quantiles for ERF, GRF and the unconditional method for U.S. wage data. In the upper half $\tau = 0.9$ and $\tau = 0.995$ holds for the lower half.

To confirm our analysis, we perform quantitative performance measures to compare ERF to the other methods. We use the prediction metric proposed by Wang and Li [2013],

$$\mathcal{R}_n(\hat{Q}.(\tau)) := \frac{\sum_{i=1}^{n} \mathbb{1}\left\{Y_i < \hat{Q}_{X_i}(\tau)\right\} - n\tau}{\sqrt{n\tau(1-\tau)}}. \tag{9}$$

In Equation (9) n is the number of observations tested and $\hat{Q}.(\tau)$ is the conditional quantile estimated on the training data set. The second half of the total sample set is partitioned into ten random folds. The different methods are fitted for each fold on the observations not used yet after which the absolute value of Equation (9) is taken. The performance of the three methods is shown in Figure 7. The 95% confidence interval is shaded grey. It is observed that ERF outperforms both GRF and the unconditional method, especially for τ closer to 1.

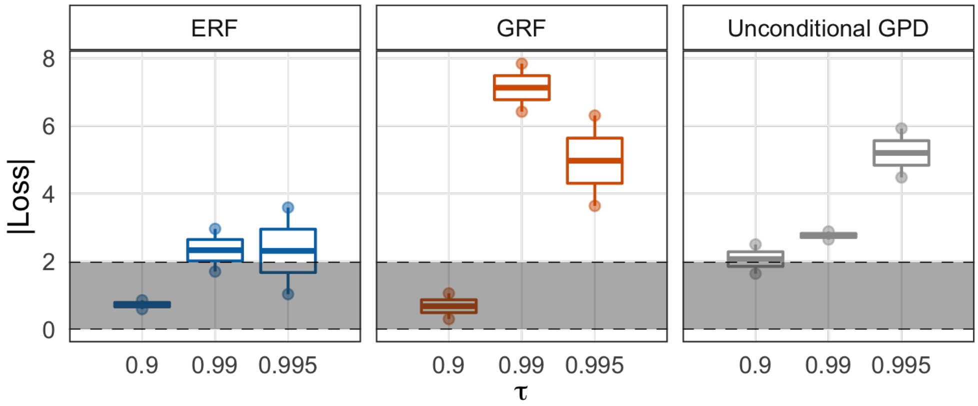

Figure 7: Absolute value of the loss (9) on the original response of the U.S. wage data. The shaded area represents the 95% interval of the absolute value of a standard normal distribution.

6.2 Analysis of Medical Claim Data

For our medical claim data, we performed the same steps as the U.S. wage data. The only difference is that we have not added 10 random predictors which would make the data set higher dimensional. Therefore, the response variable Y is the claim size, and the predictor vector consists of the numerical variable age and the binary variable whether a person is male or female. In Figure 8, the estimated GPD parameters as a function of the patient's age are shown. We observe a slightly negative correlation for the scale parameter and a positive correlation for the shape parameter. The majority observations are homogeneous between the female and male group, except for some extreme outliers for males in the scale parameter (left) graph. In this graph, the male group is formed as a saddle where the female group is more equal between different ages. The male group has a higher scale parameter for low and high ages compared to the female group. For the shape parameter, we observe a more homogeneous pattern between the female and male group. Moreover, a shift starting at age 75 for the shape parameter is observed, suggesting heavy tails for higher ages, especially for females.

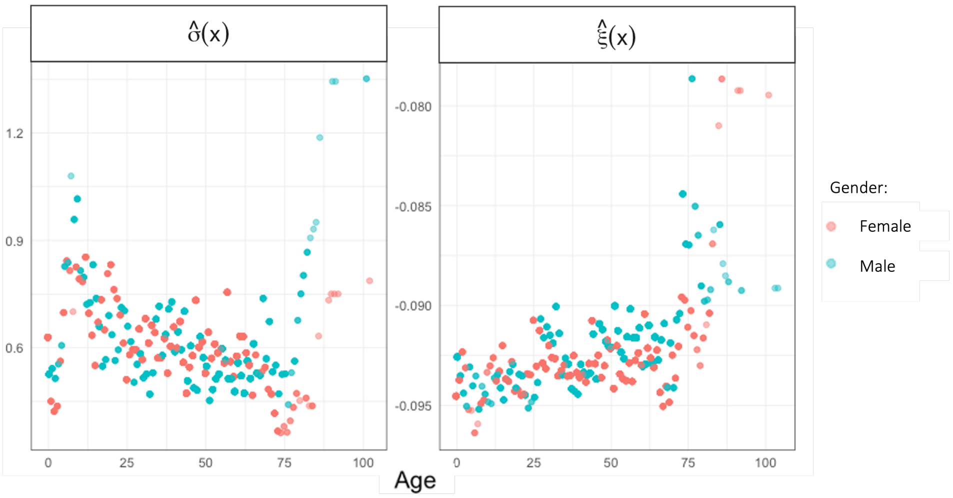

Figure 8: Estimated GPD parameters as a function of age.

Figure 9 shows the predicted quantiles for the three methods. It can be observed that the GRF gets 'out of pattern' for $\tau = 0.995$ compared to $\tau = 0.9$. This is exactly the reason we use GRF for the intermediate quantile and a method different from GRF for the extrapolation part. The unconditional method gets almost flat results for τ close to 1. This is due to the fact that following Figure 8 we expect a different scale parameter due to the shift around the age of 75. Though, the unconditional method cannot produce different scale parameters of the GPD. One cannot observe much difference in ERF between $\tau = 0.9$ and $\tau = 0.995$ suggesting the variability is modelled well.

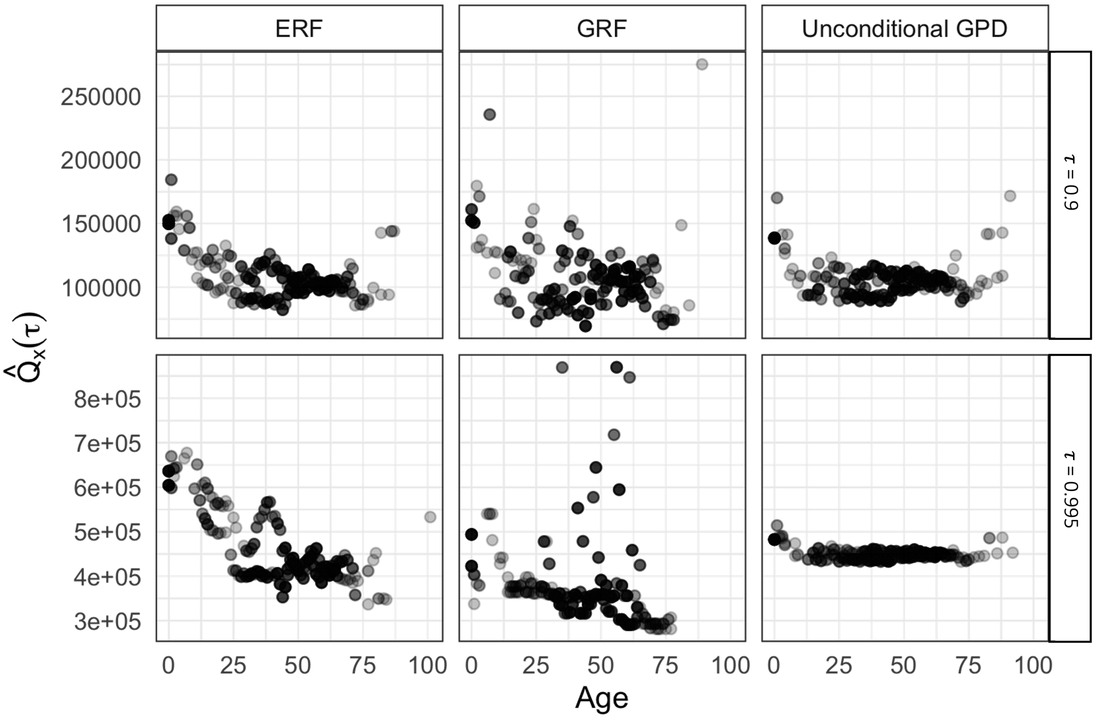

Figure 9: Predicted quantiles for ERF, GRF and the unconditional method for medical claim data. In the upper half $\tau = 0.9$ and $\tau = 0.995$ holds for the lower half.

The quantitative performance of the three methods for medical claim data is shown in Figure 10. For $\tau = 0.9$ and $\tau = 0.99$ we observe no big differences in performance between the three methods. ERF and GRF are fully in the 95% interval, but when τ gets larger and closer to 1 ($\tau = 0.995$) ERF is outperformed by both GRF and the unconditional method. As intermediate quantile method for ERF, we make use of GRF. Therefore, differences between ERF and GRF will occur within the extrapolation part. From this figure it turns out that the extrapolation of ERF for this data set is not as good as GRF nor the unconditional GPD method. Moreover, it is observed that for $\tau = 0.995$ the unconditional GPD method performs better than the ERF method. This might doubt the explanatory value of the covariates in the data set.

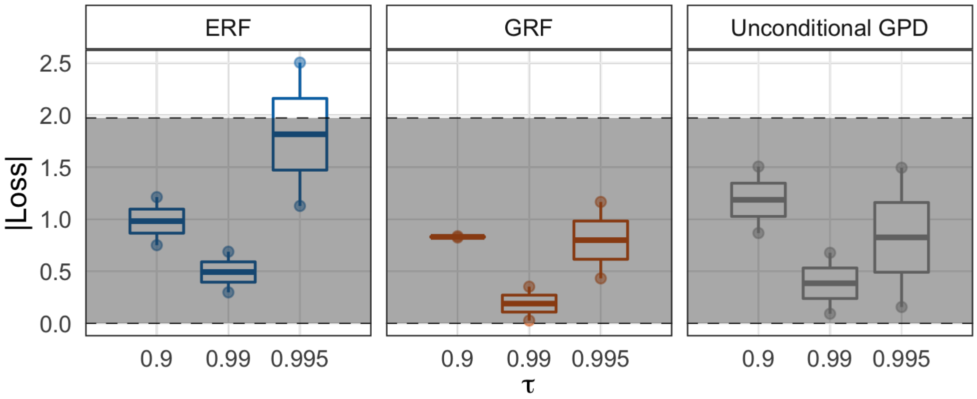

Figure 10: Absolute value of the loss (9) on the original response of the medical claim data. The shaded area represents the 95% interval of the absolute value of a standard normal distribution.

7 Conclusion and Discussion

In this thesis we studied the performance of the ERF method compared to the more classical GRF and unconditional GPD method for medical large claim data. The ERF method is proposed by Gnecco et al. [2022] and is proven to be beneficial for high dimensional data sets when estimating extreme quantiles. A high dimensional simulation study showed that ERF outperforms both the GRF and unconditional GPD method in terms of robustness and prediction performance. Besides the simulation study and the research of medical large claim data, we studied the performance of the ERF method for U.S. wage data. For the U.S. wage data it is studied what the influence of education, age, experience and whether someone is black or white is on the weekly wage.

The high dimension U.S. wage data showed good performance results for the ERF method. Whereas the GRF method has a bad performance for high quantiles. As well as for the unconditional GPD method which performs bad for high quantiles. It is shown that a jump in wage is present after 15-16 years of education, which is approximately equal to graduating a Bachelor's degree.

When investigating the medical large claim data, it is observed that performance of GRF and unconditional GPD method is better than the ERF method. One reason for this might be the not so high dimension of the predictor space for this data set. Contrary to both the simulation study as the U.S. wage application the predictor space dimension of the medical claim data is much lower. This could be a reason for the reduced performance of ERF but this has to be investigated further. An explanation for the reduced performance compared to the unconditional method could be the relevance of the covariates. It would be insightful to further investigate whether the covariates age and sex are relevant to estimating extreme quantiles for medical claim data.

Overall, this research evaluates the performance of the ERF method for estimating extreme quantiles for different data sets. By means of a simulation study it is shown that ERF outperforms the more classical GRF and unconditional GPD method. While for the U.S. wage and medical large claim data set we observe different performances of the ERF method compared to the other methods. We believe it is insightful to further investigate the relevance of the covariates within the medical large claim data.

References

J. D. Angrist, V. Chernozhukov, and I. Fernández-Val. Replication data for: Quantile Regression under Misspecification, with an Application to the U.S. Wage Structure, 2009.

S. Athey, J. Tibshirani, and S. Wager. Generalized random forests. *Annals of Statistics*, 47(2):1148–1178, 2019.

U. Balasooriya and C.-K. Low. Modeling insurance claims with extreme observations: Transformed kernel density and generalized lambda distribution. *North American Actuarial*, 12(2):129–142, 2008.

A. A. Balkema and L. de Haan. Residual Life Time at Great Age. *Annals of Probability*, 2(5):792 – 804, 1974.

L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

A. B. Casto and E. Forrestal. *Principles of Healthcare Reimbursement*. Citeseer, 2013.

A. C. Cebrián, M. Denuit, and P. Lambert. Generalized pareto fit to the society of actuaries' large claims database. *North American Actuarial Journal*, 7(3):18–36, 2003.

V. Chernozhukov. Extremal quantile regression. *Annals of Statistics*, 33(2):806 – 839, 2005.

P. de Zea Bermudez and M. Turkman. Bayesian approach to parameter estimation of the generalized pareto distribution. *Test*, 12(1):259–277, 2003.

N. Gnecco, E. Terefe, and S. Engelke. Extremal random forests. 2022.

J. H. Halton. Algorithm 247: Radical-inverse quasi-random point sequence. *Communications of the ACM*, 7(12):701–702, 1964.

T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, volume 2. Springer, 2009.

N. Meinshausen. Quantile regression forests. *Journal of Machine Learning Research*, 7:983–999, 2006.

J. Pickands. Statistical Inference Using Extreme Order Statistics. *Annals of Statistics*, 3(1):119 – 131, 1975.

R. L. Smith and J. Naylor. A comparison of maximum likelihood and bayesian estimators for the three-parameter weibull eistribution. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 36(3):358–369, 1987.

N. Tajvidi. Confidence intervals and accuracy estimation for heavy-tailed generalized Pareto distributions. *Extremes*, 6(2):111–123, 2003.

J. Uwingabire. Modelling extreme health insurance claims using generalized pareto distribution. 2018.

H. J. Wang and D. Li. Estimation of extreme conditional quantiles through power transformation. *Journal of the American Statistical Association*, 108(503):1062–1074, 2013.

T. Yaqiong. Extreme risk analysis of personal insurance claim based on block maxima method. *International Journal of Economics, Finance and Management Sciences*, 2018.