Extremal Random Forests, Tree-Based Machine Learning Methods and Extreme Value Theory For Vehicle Insurance Data

PHD THESIS

BY :
EDOSSA MERGA TEREFE

HAWASSA UNIVERSITY

HAWASSA, ETHIOPIA

APRIL, 2023

Extremal Random Forests, Tree-Based Machine Learning Methods and Extreme Value Theory For Vehicle Insurance Data

PHD THESIS

BY :

EDOSSA MERGA TEREFE

SUBMITTED TO DEPARTMENT OF STATISTICS AT HAWASSA UNIVERSITY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY IN APPLIED STATISTICS

APRIL, 2023

Approval Sheet I

This is to certify that the thesis titled " **Extremal Random Forests, Tree-Based Machine Learning Methods and Extreme Value Theory For Vehicle Insurance Data**" submitted in partial fulfillment of the requirement for the degree of **Doctor of Philosophy in Applied Statistics** to the department of Statistics, Hawassa University, and is record of original research carried out by **Edossa Merga Terefe, ID.No: PhdAps/001/10**, under my supervision and no part of the thesis has been submitted for another degree or diploma. The assistance and the help received during the course of this investigation have been duly acknowledged. Therefore, I recommended that the student has fulfilled the requirements and hence hereby can submit the thesis to the department.

Prof. Sebastian Engelke	S. Eng	23.3.23
Name of Main Supervisor	Signature	Date
Name of Co-Supervisor	Signature	Date

Approval Sheet II

We, the undersigned, members of the Board of Examiners of the final open defense by **Edossa Merga Terefe** have read and evaluated his thesis in titled **"Extremal Random Forests, Tree-Based Machine Learning Methods and Extreme Value Theory For Vehicle Insurance Data"** and Examined the candidate. This is therefore to certify that the thesis has been accepted in partial fulfillment of the requirement of the degree of Doctor of Philosophy in Applied Statistics.

Name of Department Head	Signature	Date
Prof. Sebastian Engelke	S. E.	23.3.23
Name of Main Supervisor	Signature	Date
Name of Co-Supervisor	Signature	Date
Arnoldo Frigessi	Arnoldo Frigessi	14.5.2023
Name of External Examiner	Signature	Date
Prof. Ingrid K. Glad	Ingrid K. Glad	14.05.2023
Name of internal Examiner	Signature	Date
Name of Chairperson	Signature	Date
SGS Approval	Signature	Date

Declaration

This thesis has been submitted to Department of Statistics at Hawassa University in partial fulfillment of the requirements for a degree of Doctor of Philosophy in Applied Statistics. I hereby declare that this PhD thesis is my original work and has not been submitted to any other institution and anywhere for the award of any academic degree, diploma or certificate. All sources of materials used for this thesis have been dully acknowledged.

Name of student	Signature	Date

Place: Hawassa University, Hawassa, Ethiopia

Acknowledgments

First and foremost, I would like to thank my almighty God for giving me the wisdom, strength, support and knowledge in exploring things throughout this thesis work.

It is my pleasure to express my appreciation to those who have influenced this work the most. My special gratitude directs to my main supervisor, Professor Sebastian Engelke for his incredible continuous collaboration, encouragement, guidance and kind correspondence with a full of patience up to the completion of this work. The work presented in this thesis would never have been possible without his presence. It was a real privilege and an honor for me to work with him and share some of his exceptional scientific experiences. I have been extremely lucky to have a supervisor who is extraordinary kind to me and who cares so much about not only my current work, but also about my future career. I extend my appreciation to Nicola Gnecco for his constructive suggestions and extensive discussion around my work.

I gratefully acknowledge institutions and the funding sources that made my PhD work possible. I would like to thank Wollega University, my employer for the financial support, and giving me a chance to pursue this PhD program. I appreciate to the department of Statistics at Hawassa university (Ethiopia), Oslo Centre for BioStatistics and Epidemiology at the University of Oslo (Norway) and Research Centre for Statistics at the University of Geneva (Switzerland) who took care of administrative procedures and provided me a working space and office materials during my stay at the corresponding universities.

I was funded by "NORAD" project for my one semester stay in Norway, from 01 August 2019 to 31 December 2019. During my stay in Switzerland from 01 February 2020 to 30 April 2021, I was funded by "Green Leaves Education Foundation (GLEF)" and from 01 May 2021 to 31 May 2022, the funding source was from the "Swiss National Science Foundation (SNSF)".

Furthermore, I would like to thank the Ethiopian Insurance Corporation, who gave me permission to access their vehicle insurance data set.

Finally and most importantly, my heart-felt thanks go to all members of my family for all their love, encouragement and support throughout my whole study.

I am also indebted to thank all of who participated in this thesis especially, Dr. Zeytu Gashaw, Professor Arnoldo Frigessi, my co-supervisor, Dr. Cheru Atsemegiorgis.

Thesis Outline

This thesis consists of an introduction and three research papers. The introduction part includes some background on Extreme Value Theory, Tree-Based Machine Learning Methods and Quantile Regression Forest.

Paper I:

E.M. Terefe (2023+). Tree-Based Machine Learning Methods For Vehicle Insurance Claim Prediction. https://arxiv.org/pdf/2302.10612.pdf . *Submitted for publication*

Paper II:

E.M. Terefe (2023+). Extreme Value Theory For Vehicle Insurance Data. *To be submitted for publication*

Paper III:

N. Gnecco, E.M. Terefe, and S. Engelke (2023+). Extremal Random Forests. https://arxiv.org/pdf/2201.12865.pdf. Submitted to *Journal of the American Statistical Association* and under *"Major Revision"*.

INTRODUCTION

1 Motivation

Quantile regression (QR) has received increasing attention in recent years and applied to wide areas such as investment, finance, meteorology, hydrology and economics. Compared with conventional mean regression, QR can characterizes the entire conditional distribution of the outcome variable, may be more robust to outliers and misspecification of error distribution, and provides more comprehensive statistical modeling than traditional mean regression. QR models could not only be used to detect heterogeneous effects of covariates at different quantiles of the outcome, but also offer more robust and complete estimates compared to the mean regression, when the normality assumption violated or outliers and long tails exist. These advantages make QR attractive. Existing methods for quantile regression fail if:

- The quantile of interest is extreme, i.e. close to 1, and only few or no training data points exceed it.
- The predictor space has more than a few dimensions or if the regression function of extreme quantiles is complex.

These problems motivates us to develop a method for extreme QR, which works in high dimensional data setting in Paper I.

The motivations for Paper II and Paper III are to study a novel Ethiopian vehicle insurance data, which has not been analyzed before. In Paper II, we are interested in identifying the most important predictors in predicting the claim size and analyzing their relationships using tree-based ensemble methods. The motivation for Paper III is that the insurance company has to evaluate its risk exposure due to large number of single contracts with customers. If many or very high claims have to be paid in a given period, there must be enough financial reserve to cover these losses. Thus, we create an awareness for the insurance company to give a particular care to an accurate representation of the tail of such a distribution, since the total risk will strongly depend on the amount of very high single claims.

2 Data

In this PhD study, two data sources are used: U.S. census microdata for the year 1980 (Angrist et al., 2009), which is about U.S. wage structure and Ethiopian Vehicle Insurance data. The former data set consists of 65,023 U.S.-born black and white men of age between 40-49, with five to twenty years of education, and with positive annual earnings and hours worked in the year before the census. The response Y describes the weekly wage, expressed in 1989 U.S. dollars computed as the annual income divided by the number of weeks worked. The predictor vector consists of the numerical variables age and years of education and the categorical predictor whether the person is black or white.

The second dataset is obtained from Ethiopian Insurance Corporation, one of the biggest insurance companies in Ethiopia. It consists of policy and claim information of vehicle insurance at the individual level. The dataset originally contains $n = 288,763$ unique individual contracts, represented by the observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ of $p = 10$ predictors of $\mathbf{X} = (X_1, \ldots, X_p) \in \mathbb{R}^p$ and the response variable $Y \in \mathbb{R}$ represents claim size, from July 2011 to June 2018. Some of the predictors are insured value, premium, production year of vehicle and manufacturer company.

The large number of observations in both datasets make the analysis suitable in using of tree-based machine learning methods and quantile regression at a very high quantile levels.

3 Extreme Value Theory

Historically Extreme Value Theory (EVT) analysis has been used to guide human activities against many forms of environmental disasters such as floods, storms, heat waves, winds and temperatures. However, the potential of EVT applied to the consequent financial losses from such events has only been recognized recently. EVT has recently become one of the main theories in developing statistical models for extreme losses and can be useful in defining supplementary risk measures, because it provides more appropriate distributions to fit extreme events. The heavy-tailed nature of losses from extreme events requires that special attention be put into the analysis of the tail of a loss distribution. Since a few large losses can significantly cause a huge damage to an

industry, statistical methods that deal with extreme losses have become necessary for actuaries. For example, in insurance a typical problem might be pricing or building reserves for products which offer protection against catastrophic losses, such as excess of loss reinsurance in order to price certain reinsurance treaties, which is often necessary to model losses in excess of some high threshold value, i.e., to model the largest r upper order statistics. There are two principal kinds of model for extreme values.

3.1 Block Maxima

These are models for the largest observations collected from large samples of identically distributed observations. Let $X_1, X_2, \ldots, X_n$ be independent identically distributed random variables with distribution $F(x)$. The inference is generally focused around the maximum

$$M_n = \max(X_1, X_2, \ldots X_n) \tag{3.1}$$

of the sequence. The distribution of (3.1) is easily derived by applying the rules for independent and identically distributed random variables as

$$\begin{aligned} F_{M_n}(x) &= P(X_1 < x, X_2 < x, \ldots, X_n < x) \\ &= P(X_1 < x)P(X_2 < x)\ldots P(X_n < x) = \prod_{i=1}^{n} F(x) = [F(x)]^n. \end{aligned} \tag{3.2}$$

An asymptotic approximation to $[F(x)]^n$ is based on the Fisher - Tippet theorem (1928). Given that $x < x^+$, where x^+ is the upper end-point of F (that is, the smallest value of x such that $F(x) = 1$), $[F(x)]^n \to 0$ as $n \to \infty$. The asymptotic approximation is based on the introduction of sequences of normalizing constants a_n and b_n and adjusting the distribution in (3.1) such that

$$P\left(\frac{M_n - b_n}{a_n} \leq x\right) = [F(a_n x + b_n)]^n \to G(x) \tag{3.3}$$

The Fisher and Tippet (1928) theorem states that if $G(x)$ converges to some non-degenerate

distribution function, then $G(x)$ is said to be belong to a three-parameter family, Generalized Extreme Value Distribution (GEV) of the form

$$G(x/\mu,\beta,\xi) = \begin{cases} \exp\left(-\left[1+\xi\frac{x-\mu}{\beta}\right]_+^{-\frac{1}{\xi}}\right) & \text{if } \xi \neq 0 \\ \exp\left(-\exp(-\xi\frac{x-\mu}{\beta})\right) & \text{if } \xi = 0. \end{cases} \tag{3.4}$$

where $x_+ = \max(x, 0)$.

The GEV, $G(x/\mu,\beta,\xi)$ distribution with $\beta > 0$ scale parameter, $\mu \in \mathbb{R}$ location parameter and $\xi \in \mathbb{R}$ shape parameter is defined on $\{x : 1+\xi(x-\mu)/\beta > 0\}$.

The shape parameter, ξ of the GEV distribution defines a type of distribution, meaning a family of distributions specified up to location and scaling. The GEV subsumes three types of extreme value distributions which are known by other names according to the value of ξ: when $\xi > 0$ the distribution is a Fréchet distribution; when $\xi = 0$ it is a Gumbel distribution; when $\xi < 0$ it is a Weibull distribution.

$$\text{Weibull:} \quad \Phi(x/\alpha) = \begin{cases} \exp\left[-(-x)^{\alpha}\right] & \text{if } x \leq 0 \\ 1 & \text{if } x > 0. \end{cases}, \alpha > 0$$

$$\text{Gumbel:} \quad \Lambda(x/\alpha) = \exp\left[-e^{-x}\right] \quad x \in \mathbb{R}$$

$$\text{Fréchet:} \quad \Psi(x/\alpha) = \begin{cases} 0 & \text{if } x \leq 0 \\ \exp\left[-x^{-\alpha}\right] & \text{if } x > 0. \end{cases}, \alpha > 0$$

The estimates of unknown parameters of GEV are obtained by minimizing the negative log likelihood with respect to parameter vectors (μ,β,ξ). The log negative likelihood of GEV can be written as

$$\ell(\mu,\beta,\xi) = n\log\beta + \left(1+\frac{1}{\xi}\right)\sum_{i=1}^{n}\log\left[1+\xi\left(\frac{x_i-\mu}{\beta}\right)\right] + \sum_{i=1}^{n}\left[1+\xi\left(\frac{x_i-\mu}{\beta}\right)^{\frac{-1}{\xi}}\right],$$

provided that $\left[1+\xi\left(\frac{x_i-\mu}{\beta}\right)\right] > 0 \quad \text{for } i = 1,\ldots,n.$

3.2 Peaks-Over-Threshold

The block maxima method has the major defect that it is very wasteful of data. To perform an analyses only the maximum losses in large blocks are retained. For this reason it has been largely superseded in practice by methods based on threshold exceedances, where all data that are extreme in the sense that they exceed a particular designated high level are used. Therefore, Peaks-Over-Threshold (POT) models are generally considered to be the more useful for practical applications, due to their more efficient use of the (often limited) data on extreme values in modeling the behavior of extreme values above a high threshold. An additional advantage of POT is that it provides with risk estimates that are easy to compute. Within the POT class of models one may distinguish two styles of analysis. These are the semi-parametric models built around the Hill estimator and the fully parametric models based on the generalized Pareto distribution (GPD). This thesis concentrates on the latter style of analysis for a reasons of relative simplicity in giving statistical estimates error.

The excess distribution above the threshold u can be defined as the conditional probability

$$F_u(y) = P(X-u \le y | X > u) = \frac{F(y+u)-F(u)}{1-F(u)} = \frac{F(x)-F(u)}{1-F(u)}, \qquad y > 0. \tag{3.5}$$

The methodology is based on the asymptotic approximation to the GPD of $Y = X - u$, the rescaled excesses above a suitably high level u, should the non-degenerate limiting distribution exist Pickands (1975). For those distributions F that satisfy that the distribution in (3.3) converges to (3.4), it can be shown that for large enough u there exists a positive function σ, such that (3.5) is well approximated by the the cumulative distribution function of the GPD that takes the form

$$G(y/\sigma,\xi) = \begin{cases} 1-\left(1+\frac{\xi}{\sigma}y\right)_+^{-\frac{1}{\xi}} & \text{if } \xi \neq 0 \\ 1-\exp(-\frac{1}{\sigma}y) & \text{if } \xi = 0. \end{cases} \tag{3.6}$$

where $x_+ = \max(x,0)$, defined on $\{y : y > 0\}$ and $(1+\xi y/\sigma > 0)$ with shape parameter $\xi \in \mathbb{R}$ - known as the Extreme Value Index (EVI) and threshold dependent scale parameter σ.

The GPD is applicable to very wide classes of underlying distributions of Y according to Smith (2009). The three cases $\xi < 0$, $\xi = 0$ and $\xi > 0$ correspond to different types of tail behavior. The case $\xi < 0$ arises in distributions where there is finite upper bound on the claims which are possible in generally speaking and it might be thought that this case would apply in practice. It would be expected to detect such a limit only if there is a tendency for claims to cluster near the upper limit. The second case, $\xi = 0$, which can be obtained by a formal limit $\xi \to 0$ in (3.6) typically arises in cases with an exponentially decreasing tail. This arises not only when the distribution of Y is indeed exponential, but also from many other common distributions such as gamma, Weibull, normal, lognormal etc, so it might be expected to find the estimated value of ξ close to 0 in practice. However the third case, $\xi > 0$, which is usually referred as Pareto tail is of more concern because this corresponds to a genuinely long tailed distribution and most relevant for insurance risk managers.

In the use of GPD, a critical issue in practice is the selection of an appropriate threshold u, or equivalently the selection of an adequate number of upper order statistics. There is a trade off between bias and variance in threshold selection Bawa et al. (2001). If this is set too high so that the asymptotic theorem can be considered to be essentially exact, there will not be enough data over the threshold to calculate good estimates of σ and ξ. However, it is not actually wanted u to be too low since there will not be point in basing the estimates on scores which are too small to be considered large scores, and to do so could induce a bias associated with lack of fit of the GPD.

4 Tree-Based Ensemble Methods

Ensemble methods are one of the usual choice to analyze large and complex data. Originally developed to reduce the variance-thereby improving the accuracy of an automated decision-making system, ensemble methods have since been successfully used to address a variety of statistical learning problems (Zhang and Ma, 2012), such as predictor selection, confidence estimation, error correction, among others.

The main idea behind the ensemble methodology is to weigh several individual pattern learners, and combine them in order to obtain a learner that outperforms most of them. Ensemble methodol-

ogy imitates to seek several opinions before making any crucial decision. The individual opinions are weighted, and combined to reach the final decision (Polikar, 2006).

A general principle of ensemble methods is to construct a linear combination of some model fitting method, instead of using a single fit of the method to improve the predictive performance of a given statistical learning or model fitting technique. More precisely, consider an estimation of a real-valued function $f : \mathbb{R}^p \to \mathbb{R}$ based on data $\{(x_i, y_i); i = 1, \ldots, n\}$ where x is a p-dimensional predictor variable and y a univariate response. We may then generalize to functions $f(x)$ and other data types.

Given some input data, we learn several functions $\hat{f}_1, \hat{f}_2, \hat{f}_3, \ldots, \hat{f}_B$, called learners, by changing the input data based on different reweighting. We can then construct an ensemble-based function estimate $\hat{f}_{ens}(x)$ by taking linear combinations of the individual learners as an additive expansion of the learners (Elish, 2009) $\hat{f}_i(x)$:

$$\hat{f}_{ens}(x) = \sum_{i=1}^{B} w_i \hat{f}_i(x), \tag{4.1}$$

where the $\hat{f}_i(x)$ are estimates obtained from the i^{th} reweighted dataset and w_i are the linear combination coefficients. For instance, $w_i = 1/B$, an averaging weights for Bagging (see Section 4.1) and for boosting (see Section 4.3).

4.1 Bagging

Bagging (Breiman, 1996), which stands for **b**ootstrap **agg**regat**ing**, is an ensemble method for improving unstable estimation or classification schemes. As the name implies, the two key ingredients of Bagging are bootstrap and aggregation.

Bagging adopts the bootstrap distribution for generating different learners. In other words, it applies bootstrap sampling (Efron and Tibshirani, 1993) to obtain the data subsets for training the learners. In detail, given an original data set, we generate a data set containing n number training observations by sampling with replacement. Some original observations appear more than once, while some original observations are not present in the sample. By applying the process T times, T

samples of n training observations are obtained. Then, from each sample a learner can be trained by applying the learning algorithm.

Bagging also adopts the most popular strategies for aggregating the outputs of the learners, that is, voting for classification and averaging for regression. To predict a test instance, taking regression for example, Bagging feeds the instance to its learners and collects all of their outputs, and then takes the average of the outputs as the prediction, where ties are broken arbitrarily.

In particular, the bagging method for regression is applied as follows. A learner $\hat{f}_i(x)$ is constructed in each B replica of sample, where $\hat{f}_i(x)$ is just predicted response values from $i = 1, 2, \ldots, B$ learners. Then the B learners constructed are combined using the aggregation, so that the average prediction, $f_{av}(x)$ is estimated as the average of predicted outputs from $\hat{f}_i(x)$ as:

$$\hat{f}_{av}(x) = \frac{1}{B} \sum_{i=1}^{B} \hat{f}_i(x) \tag{4.2}$$

in order to obtain a single low-variance learner.

4.2 Random Forest

Random forests (RF) (Breiman, 2001) is a non-parametric regression method that builds an ensemble model for both regression and classification trees from random subsets of features and bagged samples of the training data. It is an extension of Bagging, where the major difference with Bagging is the incorporation of randomized predictor selection. In regression settings, each decision tree predicts a test point $x \in \mathbb{R}^p$ by

$$\mu_b(x) := \sum_{i=1}^{n} \frac{\mathbb{1}\{X_i \in L_b(x)\} Y_i}{|\{i : X_i \in L_b(x)\}|}, \quad b = 1, \ldots, B,$$

where $L_b(x) \subset \mathbb{R}^p$ denotes the rectangular region that x belongs to in bth tree. By defining the similarity weights

$$w_{n,b}(x, X_i) := \frac{\mathbb{1}\{X_i \in L_b(x)\}}{|\{i : X_i \in L_b(x)\}|}, \tag{4.3}$$

the random forest predictions can be written as

$$\mu(x) := \frac{1}{B}\sum_{b=1}^{B}\mu_b(x) = \sum_{i=1}^{n} w_n(x, X_i)Y_i,$$

where $w_{n,b}(x, X_i) = \sum_{b=1}^{B} w_{n,b}(x, X_i)/B$ is the average weight across B trees.

4.3 Gradient Boosting

Gradient Boosting algorithms have been proposed in the machine learning literature by Schapire (1990) and [Freund (1995); Freund and Schapire (1996)]. The term **boosting** refers to a family of algorithms that are able to convert weak learners to strong learners for improving the predictive performance of a regression or classification procedure. Intuitively, a weak learner is just slightly better than random guess, while a strong learner is very close to perfect performance. The boosting method attempts to boost the accuracy of any given learning algorithm by fitting a series of models, each having a low error rate, and then combining them into an ensemble that may achieve better performance (Schapire, 1999). This strategy can be understood in terms of other well-known statistical approaches, such as additive models and a maximum likelihood (Friedman et al., 2000).

Like bagging, boosting is a general approach that can be applied to many ensemble statistical learner for regression or classification. Unlike bagging which is a parallel ensemble method, boosting methods are sequential ensemble algorithms where the weights w_i in (4.1) are depending on the previous fitted functions $\hat{f}_1, \ldots, \hat{f}_{i-1}$. Boosting does not involve bootstrap sampling; instead each tree is fitted on a modified version of the original data set.

There are several versions of the boosting algorithms for classification problems [Drucker (1997); Friedman et al. (2000); Schapire and Freund (2013)], but the most widely used is the one by Freund and Schapire (1996), which is known as AdaBoost, and it has been empirically demonstrated to be very accurate in terms of classification.

There are also several studies [Friedman (2002); Friedman (2001); Drucker (1997);Chen and Guestrin (2016)] conducted related to boosting for regression problems.

Boosting regression tree involves generating a sequence of trees, each grown on the residuals of

the previous tree. Therefore, boosting regression tree model inherits almost all of the advantages of tree-based models, while overcoming their primary disadvantage, that is, inaccuracy Friedman and Meulman (2003).

5 Quantile Regression Forest

Quantile regression (QR) has received increasing attention in recent years and applied to wide areas such as investment, finance, economics, medicine and engineering. Compared with conventional mean regression, QR can characterize the entire conditional distribution of the outcome variable, may be more robust to outliers and misspecification of error distribution, and provides more comprehensive statistical modeling than traditional mean regression. QR as proposed by Koenker and Bassett (1978) is a statistical technique that estimates conditional quantiles.

QR can be described as follows: let Y be a dependent variable and X a (p-dimensional) predictor variable,

$$Q_\tau(X = x) = \inf\{y : F(y \mid X = x) > \tau\}$$

The conditional distribution function $F(y \mid X = x)$ is

$$F(y \mid X = x) = P(Y \leq Y \mid X = x).$$

The objective of QR is to find the conditional quantile which minimizes the expected loss $E(\rho_\tau)$,

$$Q_\tau(X = x) = \arg\min_{q \in R} \mathbb{E}(\rho_\tau(Y - q) \mid X = x), \tag{5.1}$$

where ρ_τ is a quantile check function defined as

$$\rho_\tau(c) = c(\tau - \mathbb{I}\{c < 0\}) = \begin{cases} |c|\tau, & c \geq 0 \\ |c|(1-\tau), & c \leq 0. \end{cases}$$

We have $\rho_\tau(0) = 0$, it increases linearly with slope τ as c moves away from zero to the right and it

increases linearly with slope $1-\tau$ as c moves away from zero to the left.

Quantile regression forests (QRF), a tree-based ensemble method for estimation of conditional quantiles, has been proven to perform well in terms of prediction accuracy. QRF proposed by Meinshausen (2006) uses the conditional distribution of the response variable in the original random forest Breiman (2001) scheme. It uses the same steps as used in regression random forests to grow trees. However, at each leaf node, it retains all Y values instead of only the mean of Y values. Therefore, QRF keeps a raw distribution of Y values at each leaf node.

The expectation of the quantile loss in (5.1) cannot be estimated directly on the sample level since the set of observed predictor values does not typically include the value x. A natural estimator is

$$\hat{Q}_x(\tau) = \arg\min_{q\in\mathbb{R}} \sum_{i=1}^{n} w_n(x, X_i)\rho_\tau(Y_i - q), \tag{5.2}$$

where the set of localizing similarity weights $x' \mapsto w_n(x, x')$ are obtained according to (4.3), and it encode the similarity between the new test point x and the observed covariates X_i. In Meinshausen (2006), the weights are estimated by the standard regression random forest as (4.3) for quantile regression in (5.2). However, Athey et al. (2019) introduced generalized random forests (GRF), a method designed to fit random forests with custom loss functions. One of the main applications of GRF is quantile regression, where the trees of the forest are grown to minimize the quantile loss function.

6 Summary of the Thesis

In this thesis, one paper is dedicated to a new contribution to statistical method (Paper III). The other Papers, I and II, focus on the Ethiopian vehicle insurance data. Paper III is about quantile regression: developing a method for extreme conditional quantile estimation in a high dimensional setting and comparing the developed method with the existing literature. In this paper, we integrated extreme value theory with tree-based machine learning method. Paper I deals with claim size prediction using tree-based machine learning methods, while Paper II is about extreme value analysis in an unconditional case for Ethiopian vehicle insurance 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. URL https://doi.org/10.7910/DVN/JNEOLQ. https://doi.org/10.7910/DVN/JNEOLQ.

S. Athey, J. Tibshirani, and S. Wager. Generalized random forests. *The Annals of Statistics*, 47(2): 1148–1178, 2019. URL https://doi.org/10.1214/18-AOS1709.

J. Bawa, L. Trenner, S. Coles, and P. Dorazio. *An introduction to statistical modeling of extreme values*. Springer, 2001.

L. Breiman. Bagging predictors, 1996.

L. Breiman. Random forests. *Machine Learning, 45, 5–32*, 2001. ISSN 0885-6125. doi: 10.1023/A:1010933404324. URL http://dx.doi.org/10.1023/A:1010933404324.

T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. 2016. doi: http://dx.doi.org/10.1145/2939672.2939785.

H. Drucker. Improving regressors using boosting techniques. *In Proceedings of the 14th International Conference on Machine Learning*, 1997.

B. Efron and R. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall, New York, 1993.

M. O. Elish. Improved estimation of software project effort using multiple additive regression trees. *Expert Systems with Applications*, 36(7):10774–10778, 2009.

R. Fisher and L. Tippet. Limiting forms of the frequency distribution of the largest or smallest member of a sample. 1928.

Y. Freund. Boosting a weak learning algorithm by majority. *Information and Computation*, 121: 256–285, 1995.

Y. Freund and R. E. Schapire. Experiments with a new boosting algorithm. *In Machine Learning: Proceedings of 13th International Conference*, pages 148–156, 1996.

J. Friedman, T. Hastie, and R. Tibshirani. Additive statistical regression: a statistical view of boosting. *The Annals of Statistics*, 28(2):337–407, 2000.

J. H. Friedman. Greedy function approximation: a gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001.

J. H. Friedman. Stochastic gradient boosting. *Computational Statistics and Data Analysis*, 38(4): 367–378, 2002.

J. H. Friedman and J. J. Meulman. Multiple additive regression trees with application in epidemiology. *Statistics in Medicine*, 22(9):1365–1381, 2003.

R. Koenker and G. Bassett. Regression quantiles. *Journal of the Econometric Society*, 46(1):33–50, 1978.

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

J. I. Pickands. Statistical inference using extreme value order statistics. *Annals of Statististics*, 1975.

R. Polikar. Ensemble based systems in decision making. *Circuits and Systems Magazine, IEEE*, 6 (3):21–45, 10 2006. doi: 10.1109/MCAS.2006.1688199.

R. E. Schapire. The strength of weak learnability. *Machine Learning*, 5:197–227, 1990.

R. E. Schapire. A brief introduction to boositng. *In Proceedings of the 16th International Joint Conference on Artificial Intelligence (IJCAI'99)*, 2:1401–1406, 1999.

R. E. Schapire and Y. Freund. Boosting: Foundations and algorithms. *Kybernetes*, 42(1):164–166, 2013.

R. L. Smith. *Extreme value analysis of insurance risk*. Department of Statistics and Operations Research, University of North Carolina,Chapel Hill, NC 27599-3260, 2009.

C. Zhang and Y. Ma. *Ensemble Machine Learning: Methods and Applications*. Springer Science + Business Media, LLC, 233 Spring Street, New York, NY 10013, USA, 2012.

Tree-Based Machine Learning Methods For Vehicle Insurance Claim Size Prediction

Edossa Merga Terefe[1, 2]

[1]Research Center for Statistics, University of Geneva, Switzerland
[2]Statistics Department, Hawassa University, Ethiopia
edossa.terefe@unige.ch

Abstract

Vehicle insurance claims size prediction needs methods to efficiently handle these claims. Machine learning (ML) is one of the methods that solve this problem. Tree-based ensemble learning algorithms are highly effective and widely used ML methods. This study considers how vehicle insurance providers incorporate ML methods in their companies and explores how the models can be applied to insurance big data. We utilize various tree-based ML methods, such as bagging, random forest and gradient boosting, to determine the relative importance of predictors in predicting claims size and to explore the relationships between claims size and predictors. Furthermore, we evaluate and compare these models' performances. The results show that tree-based ensemble methods are better than the classical least square method.

Keywords: claim size prediction; machine learning; tree-based ensemble methods; vehicle insurance.

1 Introduction

A key challenge for the insurance industry is to charge each customer an appropriate price for the risk they represent. Risk varies widely from customer to customer, and a deep understanding of different risk factors helps predict the likelihood and cost of insurance claims. Thus, insurance companies must have an insurance premium that is appropriate for each customer. There are two groups in the insurance industry: life insurance and non-life insurance. This study considers nonlife insurance, particularly vehicle insurance. Insurance claims occur when the policyholder creates a formal request to an insurer for coverage or compensation for an unfortunate event of an accident. Policyholders can mitigate the costs involved with coverage for the property (damage or theft to a car) and liability (legal responsibility to others for the medical or property costs).

Insurance companies must predict how many claims are going to occur and the severity of these claims to enable insurers to set a fair price for their insurance products accordingly. In other words, claim prediction in the vehicle insurance sector is the cornerstone of premium estimates. Furthermore, it is crucial in the insurance sector to plan the correct insurance policy for each prospective policyholder.

Several studies have been done to personalize the premium estimate, such as Guillen et al. (2019) and Roel et al. (2017), they demonstrated the possible benefits of analyzing information from telematics when determining premiums for vehicle insurance. The predictive capacity of covariates obtained from telematics vehicle driving data was investigated by Gao and Wuthrich (2018) and Gao et al. (2019) using the speed–acceleration heat maps suggested by Wuthrich (2017).

Prediction accuracy enables the insurance industry to better adjust its premiums, and makes vehicle insurance coverage more affordable for more drivers. Currently, many insurance companies are transitioning to ML techniques to predict claims size. However, selecting a suitable ML predictive model is far from trivial. In this study, we investigate flexible ML techniques to make accurate predictions for claims size by analyzing a large vehicle dataset given by Ethiopian Insurance Company, one of the main car insurance company based in Ethiopia, and we apply the tree-based ML methods to the dataset, such as bagging, random forest, and gradient boosting. We also evaluate and compare the performance of these models.

The rest of this paper is organized as follows. In Section 2, the dataset is described and some descriptive statistics are provided. In Section 3, we present review of three tree-based ensemble methods is presented. In Section 4, we report the results from application of considered methods. In Section 5 we provide a discussion and conclusion of the study.

2 Dataset and Exploratory Analysis

2.1 The data

The data used for this analysis were provided from a large database of the Ethiopian Insurance Corporation, one of the biggest insurance companies in Ethiopia. It consists of policy and claim information of vehicle insurance at the individual level. The dataset originally contains $n = 288,763$ unique individual contracts, represented by the observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ where $X = (X_1, \ldots, X_p) \in \mathbb{R}^p$ denotes a vector of $p = 10$ predictors, and $Y \in \mathbb{R}$ denotes the response variable representing the claim size. The data were correspond to the period between July 2011 to June 2018. The different predictors used in the analysis are summarized in Table 2.1.

The terms, liability and comprehensive `coverage` in Table 2.1 are defined as:

- Comprehensive `coverage`: The company covers all the losses which happen to the car whenever the conditions of agreement are satisfied.

- Liability or third party `coverage`: The car can cause a damage to someone or someone's property. If the policy holder already has an agreement for a liability coverage, the insurance company covers the costs in this case. Liability amount, which a policy holder has to pay as a part of premium is nationally fixed almost every year for each car type across all the

S.N	Name	Type	Domain / Levels	Description / representation
1	Sex	categorical	0, 1, 2	0 = legal entity, 1 = male, 2 = female
2	Season	categorical	autumn, winter, spring, summer	Beginning of contract.
3	Insurance type	categorical	1201, 1202, 1204	1201 = private, 1202 = commercial, 1204 = motor trade road risk
4	Type vehicle	categorical	pick-up, truck, bus, ...	Type of vehicle grouped into six categories.
5	Usage	categorical	fare paying passengers, taxi, general cartage, ...	A usual usage of the vehicle grouped into six categories.
6	Make	categorical	Toyota, Isuzu, Nissan,...	Manufacturer company.
7	Coverage	categorical	comprehensive, liability	Scope of the insurance.
8	Production year	Integer	1960 - 2018	Vehicle's production year.
9	Insured value	continuous	$\mathbb{R}+$	Vehicle's price in USD.
10	Premium	continuous	$\mathbb{R}+$	Premium amount in USD.

Table 2.1: Description of predictors in Ethiopian vehicle insurance dataset.

insurance companies operating in the country. The liability cases are usually taken to courts or settled by negotiation. However, the affected party should not be either a family member or a relative of the policy holder.

Computation of premium is determined as a function of:

- Insured value.
- Production year: For the first three years, age of the car is not considered, but after three years, age loader computation technique, which takes into account age of the vehicle is applied.
- No claim discount (NCD): Upon the renewal of the contract after a year, a policy holder gets 20% discount from the previous year's premium and adjusted for inflation if he/she has not applied for a claim in that year. He/she can get up to 60% of discount in the consecutive years but pooled down by the age loader of the vehicle.
- Contingency, Plant and Machinery (CPM): Applicable for those cars which are operated on different circumstances. For instance, loaders type vehicles premium can be computed depending on the assumption of being at the engineering (construction) sites. But in case the vehicle cause an accident while being driven on the road, the computation needs an additional consideration and computation mechanism differs.

Some predictors such as carrying capacity and seat number are removed from the dataset prior to data analysis and modeling since they are not correctly coded.

2.2 Claims size Variable

In our analysis, the claim size is a continuous response variable $Y \in \mathbb{R}$. It is originally the amount in the Ethiopian currency Birr and it is converted to USD during data analysis. The distribution of the response variable Y is strongly zero-inflated since for about 92.5% of the contracts there is no claim paid. Thus, instead of visualizing and modeling this distribution directly, we first select the non zero observations. Given that the claim has to be paid for policy holder i, it is determined as

$$Y_i = \text{Claims Size} = \frac{\text{Insured value}_i}{\text{Market value}_i} \times \text{Loss}_i, \tag{2.1}$$

where

- Market value is the market price of the car when it was bought. The market price for each car type is collected by the insurer almost every year. The data on the market value are taken either from importers or informally from other institutions. It can be either more or less than the insured value. Knowing market value of the car helps to adjust the claim amount not to be too high in case insured value is not reliable. In most cases, insured value and market value are the same.

- Loss: When an accident happens, the damaged vehicle is inspected by the engineer experts who work in the insurance company. These experts known as server decision look at each part of the vehicle, identify the affected parts and propose either to replace or fix the vehicle. Once the affected parts have been identified, its price is determined by the server decision and a bid for repairing the damaged car is published. Including the server decision members, anyone who has a license to do so can usually participate in the bid competition. The loss is determined as follows.

In some cases, the amount of claim paid can be higher than either the insured value or market price. It is mandatory to have at least a liability insurance coverage for all vehicles as a country's regulation, even if comprehensive insurance coverage in for a safety of the vehicle's owner. Additionally, a policy holder can also have BSG and PLL insurance coverage, but if only the comprehensive coverage is already secured first. The terms BSG and PLL are defined as:

- Bandit, Shifta and Gorilla (BSG): A contract agreement in case the car is robbed or stolen. To the maximum of the insurer's knowledge, it is a vehicle insurance component applicable in Ethiopia only.
- Passengers legal liability (PLL): This is applicable for fare paying passengers, in case someone is affected by an accident being in the car. Similar to that of liability coverage, its amount is fixed to be paid as part of premium and a maximum of 40,000 birr would be paid by the insurer to a passenger in case an accident happened.

Even though both BSG and PLL insurance coverage depend on the interest of policy holder and they are optional, applicable if and only if the comprehensive insurance agreement is to take place or has already taken place.

The insured value in (2.1) does not include the values of liability, BSG and PLL, even though they are included in the contract. It contains the value of comprehensive coverage only. Thus, claims size can be higher than the insured value if:

- (total) loss + (liability insurance) + (PLL) > insured value,

where total loss is an overall loss of the car due to a severe accident and impossible to repair. In that case the insurer company pays exactly the insured value as a claim.

2.3 Exploratory Data Analysis

To make assumptions about the data and find a model that fits it best, it is important to carry out an exploratory data analysis (EDA), since it has a significant role to let the data speak for themselves prior to or as part of a formal analysis. It allows the researcher to influence and criticize an intended analysis. Additionally, EDA techniques may reveal additional information that may not be directly related to the research question. For example, EDA could suggest fruitful new lines of research (Maindonald and Braun, 2010).

The purpose of statistical graphics is to provide visual representations of quantitative and qualitative information. As a methodological tool, statistical graphics comprise a set of strategies and techniques that provide the researchers with important insights about the data under examination and help guide for the subsequent steps of the research process. The objectives of graphical methods are to explore and summarize the contents of large and complicated data sets, address questions about the variables in an analysis (for example, the distributional shapes, ranges, typical values and unusual observations), reveal structure and pattern in the data, check assumptions in statistical models, and facilitate greater interaction between the researcher and the data. Various graphical methods were examined to visualize data in raw and amalgamated formats.

The most widely recognized graphical tool to display and examine the frequency distribution and a density of a single continuous variable is the histogram.

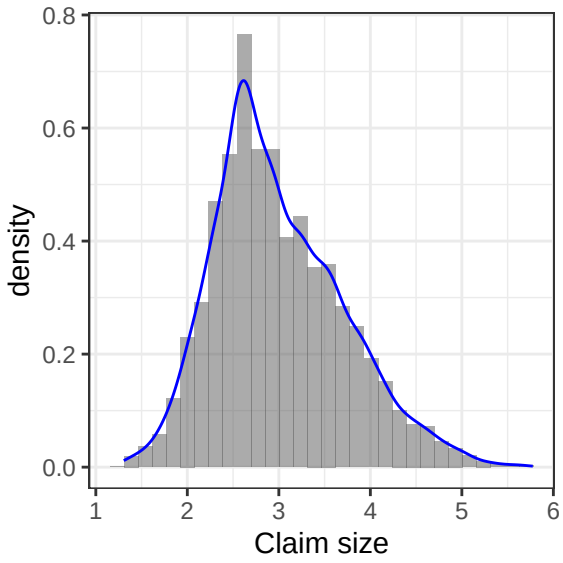

Figure 1: Frequency histogram and superimposed density plot representations of natural logarithm of claim paid distribution. The distribution of the response variable Y is strongly zero-inflated since for only about 7.5% of the contracts there is non-zero claim payment. Thus in our analysis, instead of visualizing and modeling this distribution directly, we only consider a data of policy-holders who have ever received a positive claim.

Another common tool to visualize the observed distribution of data is by plotting a smoothed histogram commonly referred as empirical density, the blue curve superimposed on the histogram with blue line in Figure 1. The empirical densities overcome some of the disadvantages caused by the arbitrary discrete bins used in the basic histograms.

2.4 Exploring relationships between covariates and Claim paid

Relationships between the predictors and the response variable can be depicted by graphical methods. Side-by-side boxplots are one way of graphical displaying the relationship between qualitative and quantitative variables. It is an excellent tool for conveying location and variation information in data sets, particularly for detecting and illustrating location and variation changes between different groups of data.

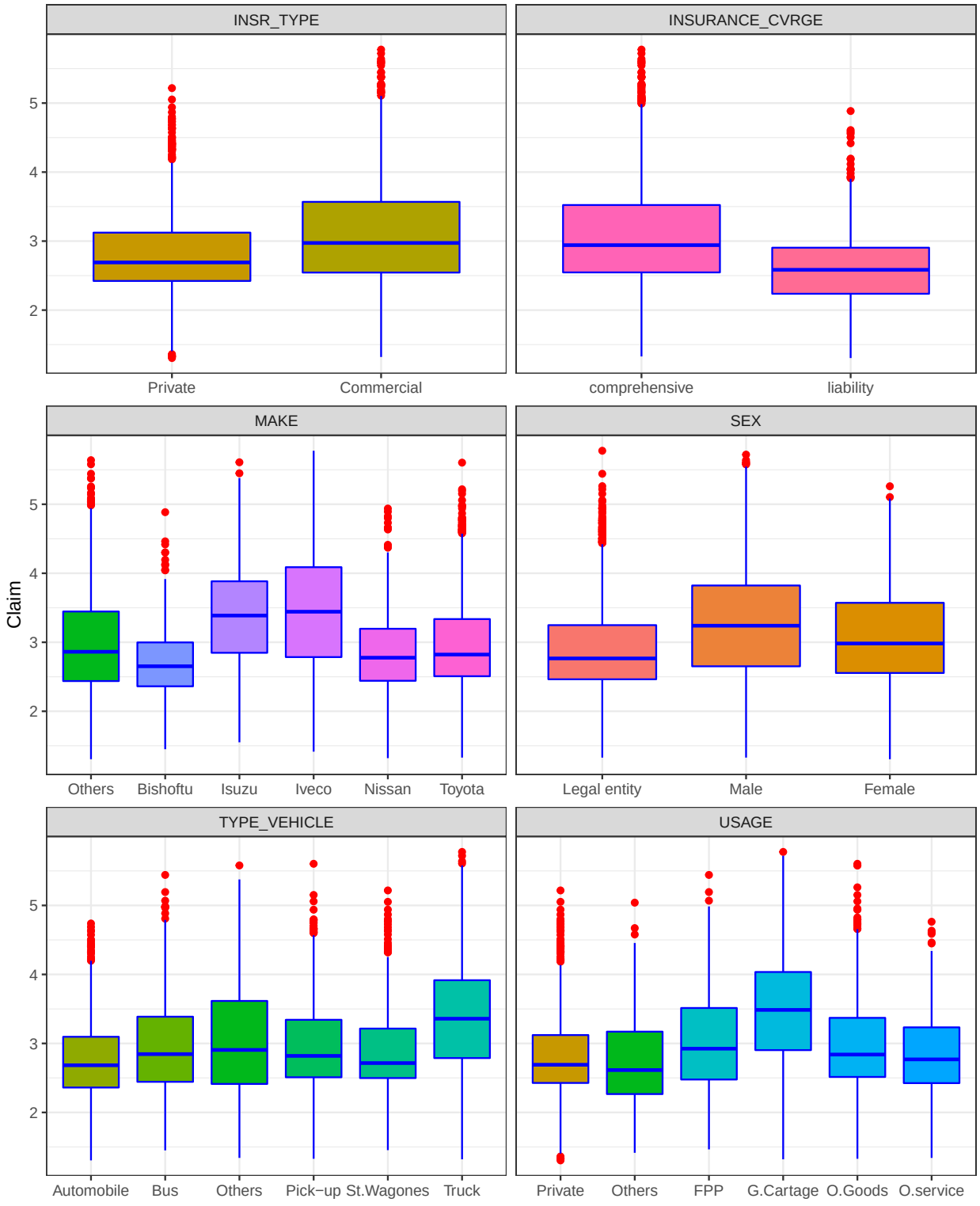

Figure 2: Boxplots of natural logarithm of claim paid against qualitative predictors.

The boxplots of claim paid against the different qualitative predictors are shown in Figure 2. Several predictors seem to have significant heterogeneity across theirs labels. For instance, in the boxplots of log of claim paid against sex, it seems existence of differences in terms of log of claim paid across the three groups of sex. Accordingly, male policy holders appeared to have a higher claim payment than either female counterparts or legal entities. Similarly, differences in claims size are observed in vehicle usage, identifying the vehicles that used for a general cartage to have the highest claim payment followed by a fare paying passengers vehicles. Moreover, vehicles manufactured from Isuzu and Iveco companies cost the insurer more than vehicles that are from any other companies, which is consistent with the insurer's prior identification of risky vehicles. It can also be seen that there are differences across the groups in other covariates such as vehicle type, insurance type and insurance coverage.

Boxplots are also a robust measure of the spread of a distribution and more informative than merely computing the variance by group as they can be helpful in identifying the homogeneity of variance between groups of a predictor. Looking at the boxplot of sex covariate again, it can be seen that the claim payment made for male policy holders appears to be more variable than either of the other two categories. And also in vehicle usage covariate, claim payment made for vehicles that are used for a general cartage and fare paying passengers purposes have more variability than any other groups. Similarly, heterogeneity of variance between a groups of insurance coverage, insurance type, manufacturer company and vehicle type covariates was observed.

Analogous to boxplots, Scatterplots are an obvious way to visualize a relationship between two quantitative variables that are numerically comparable. They are useful as a preliminary step before performing a regression analysis.

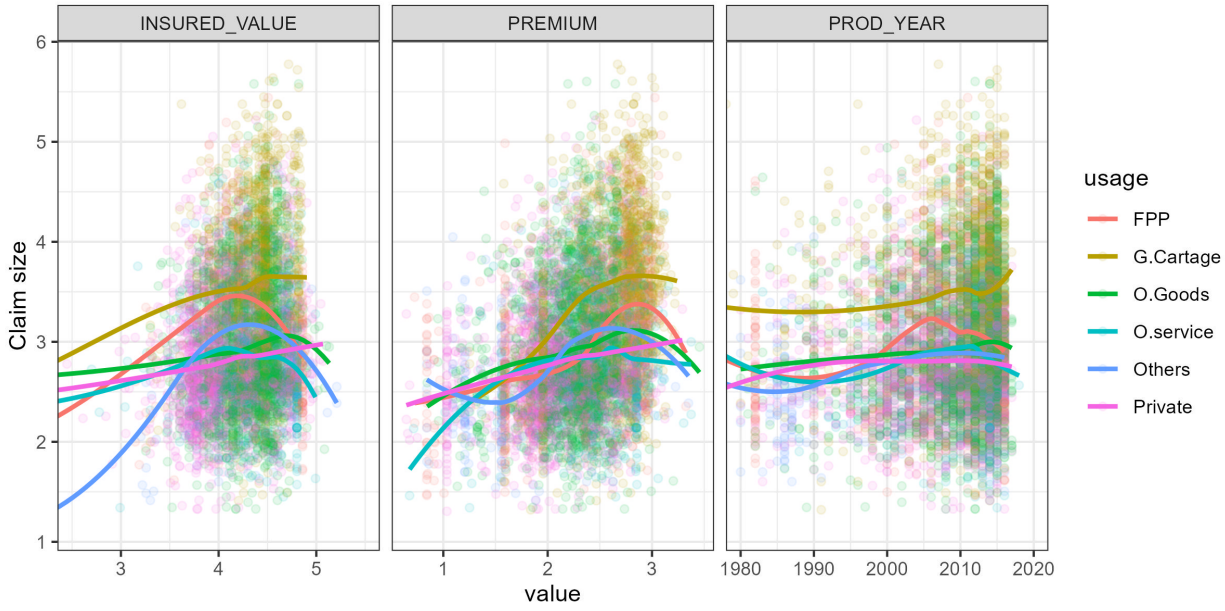

Figure 3: Scatterplots matrix: a bivariate profiling of relationships between claim and quantitative predictors.

Figure 3 shows that scatterplots, a bivariate relationships between claim paid and quantitative predictors. It is difficult to detect clear trends in any of the plots. However, by stratifying the points according to the different groups of usage predictor, we see some differences in claims size across the groups.

In addition to the scatterplot matrix seen in Figure 3, we computed correlation coefficients between claims size and insured value, premium and production year as 0.22, 0.33 and 0.11, respectively. Even though none of the coefficients between claims size and the covariates are considered to be strong, but there are some notable associations. For instance, claims size appear to have a moderate positive correlation between insured value, premium and production year , meaning that as vehicles' insured value, premium and production year increase, their claim payment also tends to increase. Correlation coefficient based relationships usually be teased out more clearly when building the (final) model.

The *loess curves* drawn on top of the scatterplots indicates a possibly nonlinear relationship between the two variables. The curves for claims size against insured value and premium are an upside-down U-shape, peaking around the middle of insured value and premium for the most groups of usage predictor. This means that the vehicles with moderate insured value and/or moderate premium have larger claims sizes than those with lower and higher insured value and /or premium. Because this trend is non-linear, this finding could not have been inferred from the correlations alone. On the other hand, when we consider the private group of usage predictor, the relationships between claim size against insured value and premium seem to be linear with a positive slope.

3 Review of Machine Learning Methods

Machine learning is now well established in many areas. In contrast to the statistical modeling approach, ML algorithms do not assume any specific model structure for the data. ML methods capture the underlying structure of data and therefore, they are more efficient in handling large data with arbitrary degree of complexity. One major task of machine learning is to construct good models from data sets.

Among ML algorithms, ensemble methods are one of the usual choice to analyze large and complex data. Originally developed to reduce the variance-thereby improving the accuracy of an automated decision-making system, ensemble methods have since been successfully used to address a variety of machine learning problems (Zhang and Ma, 2012), such as predictor selection, class imbalanced data, confidence estimation, error correction, among others.

The main idea behind the ensemble methodology is to weigh several individual pattern learners, and combine them in order to obtain a learner that outperforms most of them. In fact, combining the learners outputs does not necessarily lead to a performance that is guaranteed to be better than the best learner in the ensemble. Rather, it reduces likelihood of choosing a learner with a poor performance. Ensemble methodology imitates to seek several opinions before making any crucial decision. The individual opinions are weighted, and combined to reach the final decision (Polikar, 2006).

A general principle of ensemble methods is to construct a linear combination of some model

fitting method, instead of using a single fit of the method to improve the predictive performance of a given statistical learning or model fitting technique. More precisely, consider an estimation of a real-valued function $f : \mathbb{R}^p \rightarrow \mathbb{R}$ based on data $\{(x_i, y_i); i = 1, \ldots, n\}$ where x is a p-dimensional predictor variable and y a univariate response. We may then generalize to functions $f(x)$ and other data types.

Given some input data, we learn several functions $\hat{f}_1, \hat{f}_2, \hat{f}_3, \ldots, \hat{f}_B$, called learners, by changing the input data based on different reweighting. We can then construct an ensemble-based function estimate $\hat{f}_{ens}(x)$ by taking linear combinations of the individual learners as an additive expansion of the learners (Elish, 2009) $\hat{f}_i(x)$:

$$\hat{f}_{ens}(x) = \sum_{i=1}^{B} w_i \hat{f}_i(x), \tag{3.1}$$

where the $\hat{f}_i(x)$ are estimates obtained from the i^{th} reweighted dataset and w_i are the linear combination coefficients. For instance, $w_i = 1/B$, an averaging weights for bagging (see Section 3.1) and for boosting (see Section 3.3).

In this study, three ensemble learning algorithms i.e., bagging, random forest and boosting are considered. For models performance comparison purpose, a non-ensemble learning technique i.e., ordinary linear regression is also applied to Ethiopian vehicle insurance data set to predict claims size.

3.1 Bagging

Bagging (Breiman, 1996), which stands for **b**ootstrap **agg**regat**ing**, is an ensemble method for improving unstable estimation or classification schemes. As the name implies, the two key ingredients of Bagging are bootstrap and aggregation.

Bagging adopts the bootstrap distribution for generating different learners. In other words, it applies bootstrap sampling (Efron and Tibshirani, 1993) to obtain the data subsets for training the learners. In detail, given an original data set, we generate a data set containing n number training observations by sampling with replacement. Some original observations appear more than once, while some original observations are not present in the sample. By applying the process B times, B samples of n training observations are obtained. Then, from each sample a learner can be trained by applying the learning algorithm.

Bagging also adopts the most popular strategies for aggregating the outputs of the learners, that is, voting for classification and averaging for regression. To predict a test instance, taking regression for example, Bagging feeds the instance to its learners and collects all of their outputs, and then takes the average of the outputs as the prediction, where ties are broken arbitrarily.

In particular, the bagging method for regression is applied as follows. A learner $\hat{f}_i(x)$ is fitted on each of the B bootstrapped sample, where $\hat{f}_i(x)$ denotes the predicted response values from $i = 1, 2, \ldots, B$ learners. Then the B learners constructed are combined using the aggregation, so that the average prediction, $f_{av}(x)$ is estimated as the average of predicted outputs from $\hat{f}_i(x)$ as:

$$\hat{f}_{av}(x) = \frac{1}{B}\sum_{i=1}^{B}\hat{f}_i(x) \tag{3.2}$$

in order to obtain a single low-variance learner.

3.2 Random Forest

Random Forest (RF) is a representative of the state-of the-art ensemble methods algorithm developed by Breiman and Cutler (2008), and a very powerful technique which is used frequently in the data science field across industries (Dangeti, 2017). It is an extension of Bagging, where the major difference with Bagging is the incorporation of randomized predictor selection. During the construction of a component decision tree, at each split, RF first randomly selects a subset of predictors, and then carries out the conventional split selection procedure within the selected predictor subset.

RF is usually applied to reduce the variance of individual trees by growing numerous trees. Each subsequent split for all trees grown is not done on the entire data set, but only on the portion of the prior split that it falls under. For each tree grown, about one third of training samples are not selected in bootstrap and it is called out of bootstrap (”out of bag” or ”OOB”) samples, as in case of Bagging in Section 3.1. Using OOB samples as input to the corresponding tree, predictions are made as if they were novel test samples. A particular observation can fall in the terminal nodes of many trees in the forest, each of which, potentially, can give a different prediction. Again, the OOB sample data used to fit a particular tree is used to make each tree's prediction. Through book-keeping principle, an average for continuous response is computed for all OOB samples from all trees for the prediction of RF model. For discrete outcomes, the prediction is the majority votes from all trees that have been grown without the respective observation or the average of the predicted probabilities (Jones and Linder, 2015).

3.3 Gradient Boosting

Gradient Boosting algorithms have been proposed in the machine learning literature by Schapire (1990) and [Freund (1995); Freund and Schapire (1996)]. The term **boosting** refers to a family of algorithms that are able to convert weak learners to strong learners for improving the predictive performance of a regression or classification procedure. Intuitively, a weak learner is just slightly better than random guess, while a strong learner is very close to perfect performance. The boosting method attempts to boost the accuracy of any given learning algorithm by fitting a series of models, each having a low error rate, and then combining them into an ensemble that may achieve better performance (Schapire, 1999). This strategy can be understood in terms of other well-known statistical approaches, such as additive models and a maximum likelihood (Friedman et al., 2000).

Like bagging, boosting is a general approach that can be applied to many ensemble statistical learner for regression or classification. Unlike bagging which is a parallel ensemble method, boosting methods are sequential ensemble algorithms where the weights w_i in (3.1) are depending

on the previous fitted functions $\hat{f}_1, \ldots, \hat{f}_{i-1}$. Boosting does not involve bootstrap sampling; instead each tree is fitted on a modified version of the original data set.

There are several versions of the boosting algorithms for classification problems [Drucker (1997); Friedman et al. (2000); Schapire and Freund (2013)], but the most widely used is the one by Freund and Schapire (1996), which is known as AdaBoost, and it has been empirically demonstrated to be very accurate in terms of classification.

There are also several studies [Friedman (2002); Friedman (2001); Drucker (1997)] conducted related to boosting for regression problems. In this paper, we rely on a recently proposed gradient boosting algorithm by Chen and Guestrin (2016), which uses regression trees as the basis functions, and it optimizes a regularized learning objective function

$$\begin{aligned} L(\phi) &= \sum_i l(\hat{y}_i, y_i) + \sum_b \Omega(f_b) \\ \text{where} \quad \Omega &= \gamma T + \frac{1}{2}\lambda ||w||^2. \end{aligned} \tag{3.3}$$

Here, l is a differentiable convex loss function that measures the difference between the prediction $\hat{y}_i$ and the target y_i. The second term Ω penalizes the complexity of the model (i.e., the regression tree functions). The additional regularization term helps to smooth the final learned weights to avoid over-fitting.

Boosting regression tree involves generating a sequence of trees, each grown on the residuals of the previous tree. Therefore, boosting regression tree model inherits almost all of the advantages of tree-based models, while overcoming their primary disadvantage, that is, inaccuracy Friedman and Meulman (2003).

4 Application of Machine Learning Methods to Insurance Data

4.1 Variable Importance

Our goal is not only to find the most accurate model of the response, but also to identify which of the predictor variables are most important to make the predictions. For this reason, we perform variable importance. The ensemble methods algorithm estimate the importance of for instance, x_j predictor variable by looking at how much prediction error increases over the baseline error, when the OOB sample for x_j predictor is permuted while all others are left unchanged. The most commonly used variable importance measure is the permutation importance, introduced by Breiman (2001), which suggests that the variable importance of predictor x_j is the difference in prediction accuracy after and before permuting x_j averaged over all trees. More precisely, the variable importance of predictor x_j is defined as

$$\text{VI}(x_j) = \frac{1}{B}\sum_{i=1}^{B} \text{VI}\left(x_j\right)_i \tag{4.1}$$

where $\text{VI}\left(x_j\right)_i = \left(\text{RMSE}_i^{x_j} - \text{RMSE}_i^0\right)$, and $\text{RMSE}_i^{x_j}$ representing the RMSE value for the i^{th}

model in the ensemble fitted from the dataset with a random permutation applied to the covariate x_j, and RMSE_i^0 is the RMSE value for this model fitted to the original dataset. Note that $\text{VI}\left(x_j\right)_i = 0$, if variable x_j is not in the i^{th} model. The raw variable importance score for each variable is then computed as the mean importance over all trees. In fact, $\text{VI}(x_j)$ can be computed depending on any other performance measure such as coefficient of determination, R^2.

Let $x_1, \ldots, x_p$ be the features of interest and let RMSE_0 be the baseline performance metric for the trained model. The permutation-based variable importance scores can be computed as shown in Algorithm 1.

Algorithm 1 Permutation-based variable importance computation.

1: **For** $j \in \{1, 2, \ldots, p\}$;

   (a) Permute the values of feature x_i in the training data.

   (b) Recompute the performance metric on the permuted data, RMSE.

   (c) Record the difference from baseline using 4.1.

2: Return the variable importance scores $\text{VI}(x_1), \ldots, \text{VI}(x_p)$.

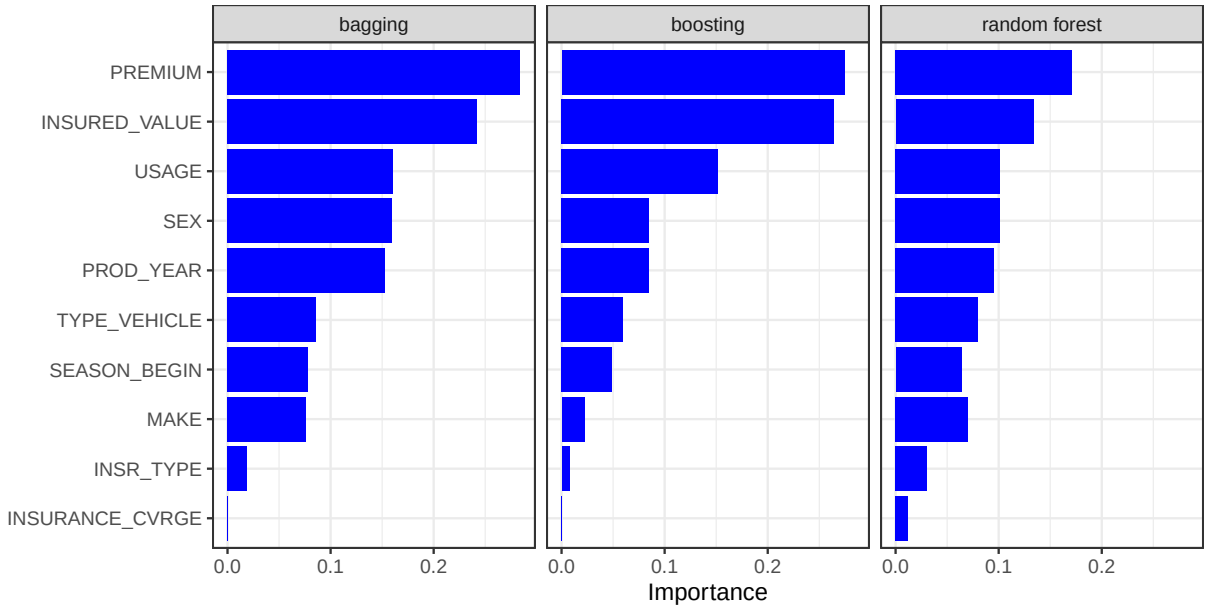

Figure 4: A graphical representation of average variable importance across all the trees from Bagging, boosting and regression RF. The larger the number, the bigger the effect.

Figure 4 displays the importance predictor while growing the trees. Accordingly in all the three models, premium is the most crucial predictor followed by the insured value. The second most

influential (slightly equally in bagging and random forest) predictors are usage and sex, followed by production year, in line with the earlier boxplots exploratory analysis.

Figure 4 is obtained by repeating the permutation of each variable 20 times and the results averaged together. This helps to provide more stable VI scores, and also the opportunity to measure their variability as seen in Figure 5, since permutation approach introduces randomness into the procedure.

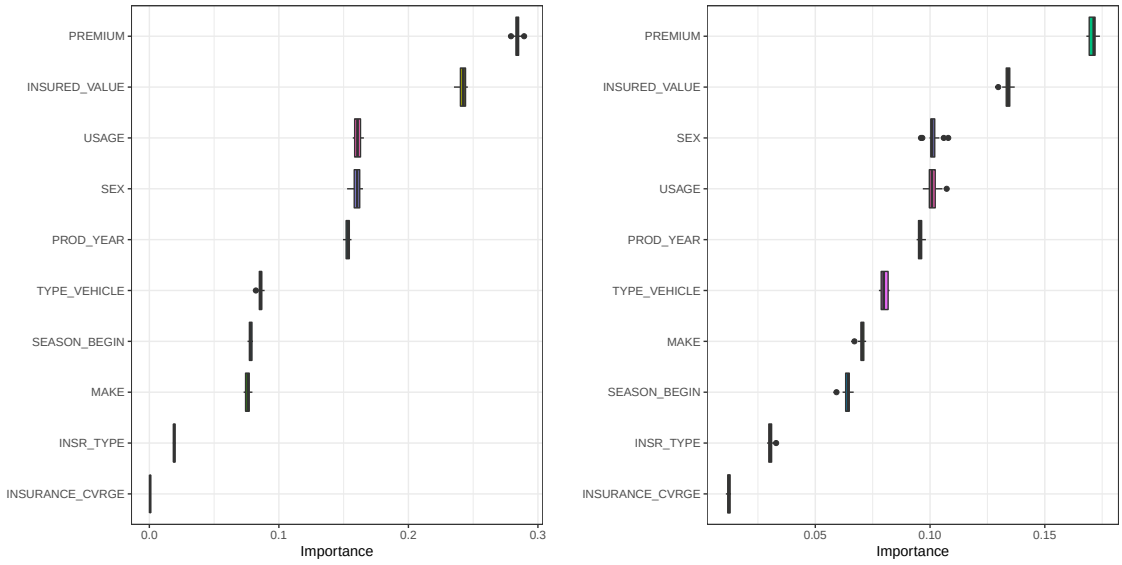

Figure 5: Boxplots of variable importance from Bagging (left panel) and regression RF (right panel) from 20 times repeated permutation.

4.2 Partial Dependence Plots

Though determining predictor importance is a crucial task in any supervised learning problem, ranking variables is only part of the story and once the important predictors are identified it is often necessary to assess the relationship between the predictors and the response variable. The task is often accomplished by constructing partial dependence plots (PDP) (Friedman, 2001), which helps to visualize the relationship between a predictor and the response variable while accounting for the average effect of the other predictors in the model.

Let $\hat{y}$ be prediction function from an arbitrary model using a dataset, $D = \left\{(x_{i,j}, y_i)\right\}$ for $i = 1, \ldots, n$ and $j = 1, \ldots, p$. The model generates predictions of the form:

$$\hat{y}_i = f\left(x_{i,1}, x_{i,2}, \ldots, x_{i,p}\right), \tag{4.2}$$

for some function $f(...)$.

Let x_k be a single predictor of interest with unique values $\left(x_{1,k}, x_{2,k}, \ldots, x_{n,k}\right)$. Then the partial dependence plots are obtained by computing the following average and plotting it over a useful range of x values:

$$\bar{f}_k(x) = \frac{1}{n}\sum_{j=1}^{n} \hat{f}\left(x_{1,j}, \ldots, x_{k-1,j}, x, x_{k+1,j}, \ldots, x_{p,j}\right) \tag{4.3}$$

The function $\bar{f}_k(x)$ indicates how the value of the variable x_k influences the model predictions $\{y_j\}$ after we have averaged out the influence of all other variables. The partial dependence of the response on x_k can be constructed as Algorithm 2:

Algorithm 2 Partial dependence construction of the response on a single predictor x_k.

1: **For** $j \in \{1, 2, \ldots, n\}$;

(a) Replace the original training values of x_k with the constant $x_{k,j}$.

(b) Compute vector of predicted values, $\{\hat{y}_j\}$ from the modified version of the training data.

(c) Compute the average prediction according to 4.3 to obtain $\bar{f}_k(x_{k,j})$.

2: Plot the pairs of $\{x_{k,j}, \bar{f}_k(x_{k,j})\}$ for $j = 1, 2, \ldots, n$

Since Algorithm 2 can be quite computationally intensive as it involves n passes over the training records, a reduced number of points is used by equally spacing the values in the range of the variable of interest.

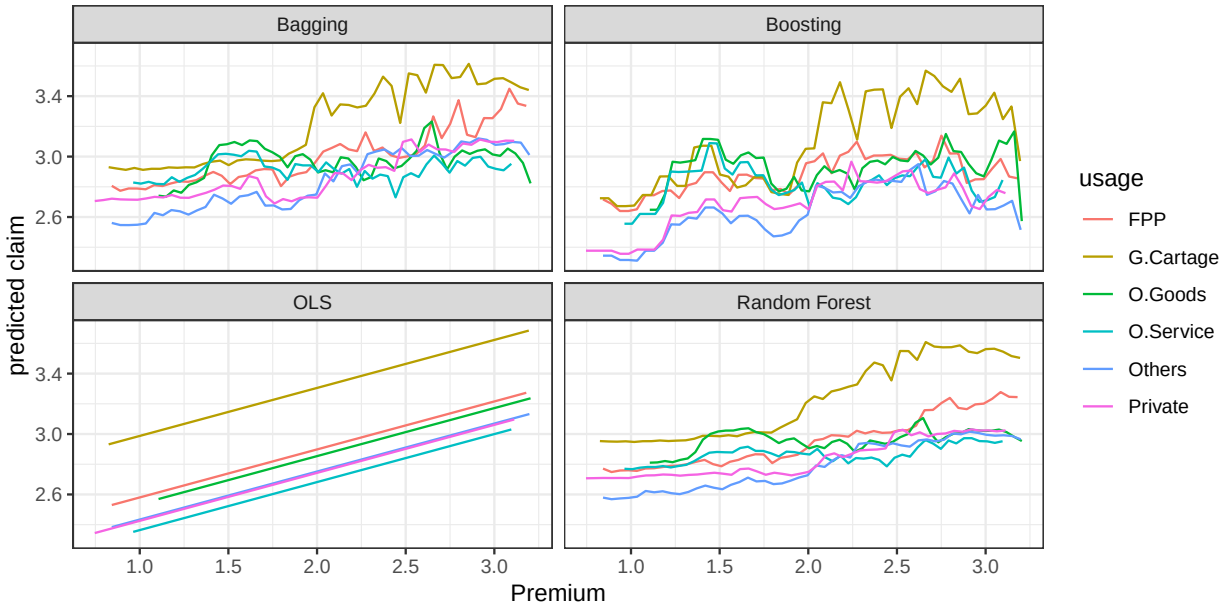

Figure 6: Two-way partial dependence plots; the marginal effect of premium on the claims size for different groups of usage predictor after integrating out the other variables.

Figure 6 shows a separate partial dependency function for each group of usage predictor. Because one-way partial dependency plots display one predictor at a time, they are valid only if

the predictor of interest does not interact strongly with other predictors. However, interactions are common in actual practice; in these cases, we can use higher-order (such as two- and three-way) partial dependence plots to check for interactions among predictors. For example, Figure 6 shows an interaction between premium and usage predictors.

The two-way plot shows that vehicles used for a general cartage with a high premium (more than 2 in ensemble methods) have much higher expected claims size compared to the vehicles with other usages. This interaction would have not apparent in the one-way plot.

4.3 Methods Comparison

In this application, the predictor variable is represented by a collection of quantitative and qualitative attributes of the vehicle and the response is the actual claims size. Given a collection of N observations $\{(x_i, y_i); i = 1, \ldots, N\}$ of known (x, y) values, the goal is to use this data to obtain an estimate of the function that maps the predictor vector x to the values of the response variable y. This function can then be used to make predictions on observations where only the x values are observed. Formally, we wish to learn a prediction function $x \mapsto \hat{f}(x)$ that minimizes the expectation of some loss function $\mathcal{L}(\hat{f}(x), y)$ over the joint distribution of all (x, y)-values, that is

$$\hat{f}(x) = arg\, min_{f(x)} E\left[\mathcal{L}(f(x), Y)|X = x\right]. \tag{4.4}$$

In finite samples, we evaluate the performance of $\hat{f}$ with the Mean Square Error (MSE), that is

$$MSE = \frac{1}{n'} \sum_{i=1}^{n'} \left(\hat{f}(x_i) - y_i\right)^2 \tag{4.5}$$

where $\hat{f}(x)$ is a fitted regression function on the test data set $\{x_i\}_{i=1}^{n'}$ and y is the observed response variable.

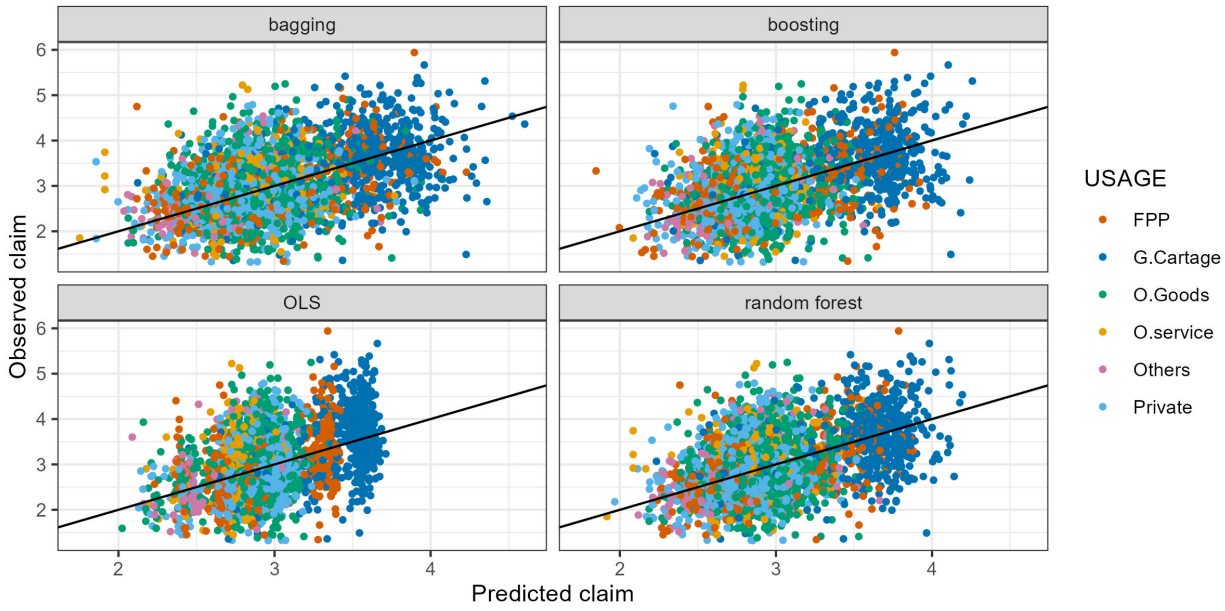

Figure 7: Observed against predicted claims size on log scales

For statistical modeling purposes, we first partitioned the data into train (70%) and test (30%) data sets. The train set was used for exploratory data analysis, model training and selection, and the test set to assess the predictive accuracy of the selected method. The training data goes from years 2011 to 2015 and from 2017 to 2018, while the test data is observations from the year 2016.

Regarding performance of the methods, it can be clearly seen that OLS method is predicting all the claims less than USD 10^4 even though some observed claims are even larger than USD 10^5. However in case of ensemble methods, they could predict claims size beyond 10^4. Both OLS and ensemble methods underestimated high claims size, but the underestimation is higher in OLS.

5 Conclusion and Discussion

Ensemble methods are well established algorithms for obtaining highly accurate classifiers by combining less accurate ones. This paper has provided a brief overview of methods for constructing ensembles and reviewed the three most popular ones, namely bagging, random forest and gradient boosting. The paper has also provided some results from application of ensembles on real vehicle insurance dataset to address some problems of insurance companies. In the application section, the predictors are ranked according to their importance in predicting claims size, and the relationships between claims size and some predictors are assessed. The performances of non-ensemble (OLS) and ensemble learning algorithms (bagging, random forests and gradient boosting) are evaluated in terms of RMSE. Accordingly, the ensemble learning techniques outperformed the OLS. Thus, this study suggests that ensemble learning techniques can outperform non-ensemble techniques. Moreover, the three ensemble algorithms performed similarly.

In this paper, we applied methods which are tailored to conditional mean prediction. These methods do not take into account the effect of single large claim sizes on the total risk of insurance companies. Thus, particular care have to be given to an accurate modeling of a very few large claim sizes, since the total risk of insurance companies will strongly depend on those claim sizes. A popular way to do this is through quantile regression (Koenker and F.Hallock, 2001) framework in a conditional extreme settings. Among several approaches available on the conditional extreme quantile estimation, [Gnecco et al. (2022); Pasche and Engelke (2022); Velthoen et al. (2021); Youngman (2019)] integrated some flexible methods such as Generalized Random Forest (GRF) (Athey et al., 2019) with the Generalized Pareto Distribution (GPD). They all model the parameters of the GPD as flexible functions. Therefore, it would be very interesting and worth to apply some versions of Extreme Value Theory (EVT) on the dataset used in this paper in future.

References

S. Athey, J. Tibshirani, and S. Wager. Generalized random forests. *The Annals of Statistics*, 47(2): 1148–1178, 2019. URL https://doi.org/10.1214/18-AOS1709.

L. Breiman. Bagging predictors, 1996.

L. Breiman. Random forests. *Machine Learning, 45, 5–32*, 2001. ISSN 0885-6125. doi: 10.1023/A:1010933404324. URL http://dx.doi.org/10.1023/A:1010933404324.

L. Breiman and A. Cutler. *Random Forests: Classification/Clustering Manual*. Statistics Department, University of California, Berkeley, CA 94720, 2008. URL http://www.math.usu.edu/~adele/forests/cc_home.htm.

T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. 2016. doi: http://dx.doi.org/10.1145/2939672.2939785.

P. Dangeti. *Statistics for machine learning: build supervised, unsupervised, and reinforcement learning models using both Python and R*. Packt Publishing Ltd, Birmingham B3 2PB, UK., 2017. ISBN 978-1-78829-575-8.

H. Drucker. Improving regressors using boosting techniques. *In Proceedings of the 14th International Conference on Machine Learning*, 1997.

B. Efron and R. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall, New York, 1993.

M. O. Elish. Improved estimation of software project effort using multiple additive regression trees. *Expert Systems with Applications*, 36(7):10774–10778, 2009.

Y. Freund. Boosting a weak learning algorithm by majority. *Information and Computation*, 121: 256–285, 1995.

Y. Freund and R. E. Schapire. Experiments with a new boosting algorithm. *In Machine Learning: Proceedings of 13th International Conference*, pages 148–156, 1996.

J. Friedman, T. Hastie, and R. Tibshirani. Additive statistical regression: a statistical view of boosting. *The Annals of Statistics*, 28(2):337–407, 2000.

J. H. Friedman. Greedy function approximation: a gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001.

J. H. Friedman. Stochastic gradient boosting. *Computational Statistics and Data Analysis*, 38(4): 367–378, 2002.

J. H. Friedman and J. J. Meulman. Multiple additive regression trees with application in epidemiology. *Statistics in Medicine*, 22(9):1365–1381, 2003.

G. Gao and M. V. Wuthrich. Feature extraction from telematics car driving heatmaps. *European Actuarial Journal*, 8:383 – 406, 2018.

G. Gao, S. Meng, and M. V. Wuthrich. Claims frequency modeling using telematics car driving data. *Scandinavian Actuarial Journal*, pages 62 – 143, 2019.

N. Gnecco, E. M. Terefe, and S. Engelke. Extremal random forests, 2022. URL https://arxiv.org/abs/2201.12865.

M. Guillen, J. P. Nielsen, M. Ayuso, and A. M. Pérez-Marín. The use of telematics devices to improve automobile insurance rates. *Risk Analysis*, 39:72 – 662, 2019.

Z. Jones and F. Linder. Exploratory data analysis using random forests. 73rd annual MPSA conference, 2015.

R. Koenker and K. F.Hallock. Quantile regression. *Journal of Economic Perspectives*, 15(4): 143–156, 2001.

J. Maindonald and W. J. Braun. *Data Analysis and Graphics Using R: An example-based approach*. Cambridge series in Statistical and Probabilistic Mathematics. Cambridge University press 2003, Mathematical Sciences Institute, Australian National University, third edition edition, 2010. URL www.cambridge.org/9780521762939.

O. C. Pasche and S. Engelke. Neural networks for extreme quantile regression with an application to forecasting of flood risk, 2022. URL https://arxiv.org/abs/2208.07590.

R. Polikar. Ensemble based systems in decision making. *Circuits and Systems Magazine, IEEE*, 6 (3):21–45, 10 2006. doi: 10.1109/MCAS.2006.1688199.

V. Roel, K. Antonio, and G. Claeskens. Unraveling the predictive power of telematics data in car insurance pricing. *Journal of the Royal Statistical Society SSRN*, 28:72 – 112, 2017.

R. E. Schapire. The strength of weak learnability. *Machine Learning*, 5:197–227, 1990.

R. E. Schapire. A brief introduction to boositng. *In Proceedings of the 16th International Joint Conference on Artificial Intelligence (IJCAI'99)*, 2:1401–1406, 1999.

R. E. Schapire and Y. Freund. Boosting: Foundations and algorithms. *Kybernetes*, 42(1):164–166, 2013.

J. Velthoen, C. Dombry, J.-J. Cai, and S. Engelke. Gradient boosting for extreme quantile regression, 2021. URL https://arxiv.org/abs/2103.00808.

M. V. Wuthrich. Covariate selection from telematics car driving data. *European Actuarial Journal*, 7:89 – 108, 2017.

B. D. Youngman. Generalized additive models for exceedances of high thresholds with an application to return level estimation for u.s. wind gusts. *Journal of the American Statistical Association*, 114(528):1865–1879, 2019. doi: 10.1080/01621459.2018.1529596. URL https://doi.org/10.1080/01621459.2018.1529596.

C. Zhang and Y. Ma. *Ensemble Machine Learning: Methods and Applications*. Springer Science + Business Media, LLC, 233 Spring Street, New York, NY 10013, USA, 2012.

Extreme Value Theory For Vehicle Insurance Data

Edossa Merga Terefe[1, 2]

[1]Research Center for Statistics, University of Geneva, Switzerland
[2]Statistics Department, Hawassa University, Ethiopia
edossa.terefe@unige.ch

Abstract

We study the Ethiopian vehicle insurance dataset using the models block maxima and peaks-over-threshold based on extreme value theory for estimating the risk measures, Value-at-Risk and Expected Shortfall. The extreme observations are fitted to the generalized extreme value distribution and the generalized Pareto distribution using maximum likelihood estimation. When estimating the model parameters and risk measures, the difference in estimates between the models is observed.

Keywords: claim size; expected shortfall; GEV; Value-at-Risk; POT.

1 Introduction

The aim of this paper is to illustrate the tail distribution estimation of a claim size for Ethiopian vehicle insurance dataset. The illustration focuses on the graphical visualization of the claim size, and providing point and confidence intervals of estimates. We considered the right tail of the claim size distribution, and it is considered as a negative returns or losses. Then we used the estimation results to quantify the risk measures. The two risk measures considered are Value at Risk (VaR) and Expected Shortfall (ES). These two measures are used to predict how much can the claim size can rise. VaR is equal to the smallest claim size such that the probability of obtaining a greater claim size, is less than or equal to some predetermined probability α. Further, ES can be summarized as the average of the claim size that are greater than VaR. Hence, when calculating VaR, a lower limit of "the worst claim size" is obtained, while when calculating ES the average of these "worst claim sizes" is produced [see McNeil et al. (2005) and Hull (2018)].

A number of models exist for computing VaR and ES. Here, we focus on two different models based on extreme value theory. Extreme value theory is used to analyze events that happen rarely, i.e., extreme events. In our setting, rare events consist of large claim sizes in the vehicle insurance dataset. The two models based on extreme value theory are called block maxima and peaks-over-threshold (POT). Both models have the same objective; fit a distribution to the sample of extreme

observations. However, the models assume that the data follow different distributions. Also, which observations from the original sample that should be considered as extreme, differs in the two models [see Coles (2001) and Dowd (2005)].

We first consider the block maxima method using GEV distribution, which allows the determination of the VaR and ES. Second, we model the exceedances over a given threshold using GPD, which enables us to estimate high quantiles of the claim paid distribution and the corresponding ES. GEV and GPD are two different distributions, but they have the same purpose: model the distribution of the extreme claim size. In particular, we can note that shape parameter, denoted ξ, is contained in both distributions, and it should therefore take similar values (and same sign) in the two distributions (Coles, 2001). In this paper, a positive ξ is obtained in both models, which is the case in Gilli and Këllezi (2006). Dowd (2005) and McNeil et al. (2005) express that the case $\xi < 0$ is often not of great interest since most of insurance data are more heavily tailed.

The remaining of this paper is summarized as follows. We start with a Section 2 of the brief review of extreme value theory and risk measures. Before the main results are given in Section 4, the data used and exploratory data analysis are presented in Section 3. We close the paper with a discussion and conclusion of the results in Section 5.

2 Background

2.1 Extreme Value Theory

Assessing the probability of extreme events in summarizing its distribution with a risk measure is an important issue mainly in managing a risk of financial portfolios, since the viability of the insurance industry depends on probabilistic calculations of risk. Extreme Value Theory (EVT) has recently become one of the main theories in developing statistical models for extreme insurance losses and can be useful in defining supplementary risk measures, because it provides more appropriate distributions to fit extreme events. The heavy-tailed nature of insurance claims requires that special attention be put into the analysis of the tail of a loss distribution. Since a few large claims can significantly impact an insurance portfolio, statistical methods that deal with extreme losses have become necessary for actuaries. For example, in insurance a typical problem might be pricing or building reserves for products which offer protection against catastrophic losses, such as excess of loss reinsurance in order to price certain reinsurance treaties, which is often necessary to model losses in excess of some high threshold value, i.e., to model the largest r upper order statistics. There are two principal kinds of model for extreme values.

2.1.1 Block Maxima

These are models for the largest observations collected from large samples of identically distributed observations. Let $X_1, X_2, \ldots, X_n$ be independent identically distributed random variables with distribution $F(x)$. The inference is generally focused around the maximum

$$M_n = \max(X_1, X_2, \ldots X_n) \tag{2.1}$$

of the sequence. The distribution of (2.1) is easily derived by applying the rules for independent and identically distributed random variables as

$$\begin{aligned} F_{M_n}(x) &= P(X_1 < x, X_2 < x, \ldots, X_n < x) \\ &= P(X_1 < x)P(X_2 < x)\ldots P(X_n < x) = \prod_{i=1}^{n} F(x) = [F(x)]^n. \end{aligned} \tag{2.2}$$

An asymptotic approximation to $[F(x)]^n$ is based on the Fisher - Tippet theorem (1928). Given that $x < x^+$, where x^+ is the upper end-point of F (that is, the smallest value of x such that $F(x) = 1$), $[F(x)]^n \to 0$ as $n \to \infty$. The asymptotic approximation is based on the introduction of sequences of normalizing constants a_n and b_n and adjusting the distribution in (2.1) such that

$$P\left(\frac{M_n - b_n}{a_n} \leq x\right) = [F(a_n x + b_n)]^n \to G(x) \tag{2.3}$$

The Fisher and Tippet (1928) theorem states that if $G(x)$ converges to some non-degenerate distribution function, then $G(x)$ is said to be belong to a three-parameter family, Generalized Extreme Value Distribution (GEV) of the form

$$G(x/\mu, \beta, \xi) = \begin{cases} \exp\left(-\left[1 + \xi\frac{x-\mu}{\beta}\right]_+^{-\frac{1}{\xi}}\right) & \text{if } \xi \neq 0 \\ \exp\left(-\exp(-\xi\frac{x-\mu}{\beta})\right) & \text{if } \xi = 0. \end{cases} \tag{2.4}$$

where $x_+ = \max(x, 0)$.

The GEV, $G(x/\mu, \beta, \xi)$ distribution with $\beta > 0$ scale parameter, $\mu \in \mathbb{R}$ location parameter and $\xi \in \mathbb{R}$ shape parameter is defined on $\{x : 1 + \xi(x - \mu)/\beta > 0\}$.

The shape parameter, ξ of the GEV distribution defines a type of distribution, meaning a family of distributions specified up to location and scaling. The GEV subsumes three types of extreme value distributions which are known by other names according to the value of ξ: when $\xi > 0$ the distribution is a Fréchet distribution; when $\xi = 0$ it is a Gumbel distribution; when $\xi < 0$ it is a Weibull distribution.

$$\text{Weibull:} \quad \Phi(x/\alpha) = \begin{cases} \exp\left[-(-x)^\alpha\right] & \text{if } x \leq 0 \\ 1 & \text{if } x > 0. \end{cases}, \alpha > 0$$

$$\text{Gumbel:} \quad \Lambda(x/\alpha) = \exp\left[-e^{-x}\right] \quad x \in \mathbb{R}$$

$$\text{Fréchet:} \quad \Psi(x/\alpha) = \begin{cases} 0 & \text{if } x \leq 0 \\ \exp\left[-x^{-\alpha}\right] & \text{if } x > 0. \end{cases}, \alpha > 0$$

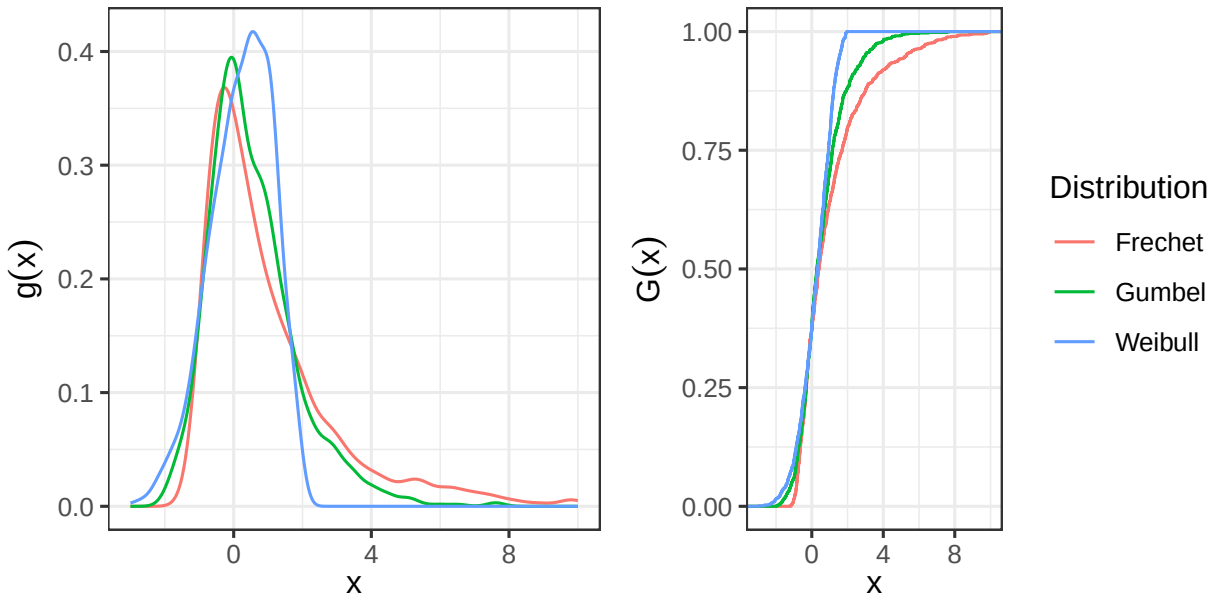

Figure 1: The density function of a standard GEV distribution in three cases: Weibull ($\xi = -0.5$); Gumbel ($\xi = 0$); and Fréchet ($\xi = 0.5$), and their corresponding distributions. In all cases $\mu = 0$ and $\beta = 1$.

The density and distribution function of the GEV distribution are shown in the left and right panels of Figure 1, respectively for the three cases $\xi = -0.5$, $\xi = 0$ and $\xi = 0.5$, corresponding to Weibull, Gumbel and Fréchet types, respectively. Observe that the Weibull distribution is a short-tailed distribution with a so-called finite right endpoint. The right endpoint of a distribution will be denoted by $x_F = \sup\{x \in \mathbb{R} : F(x) < 1\}$. The Gumbel and Fréchet distributions have infinite right endpoints, but the decay of the tail of the Fréchet distribution is much slower than that of the Gumbel distribution.

The estimates of unknown parameters of GEV are obtained by minimizing the negative log likelihood with respect to parameter vectors (μ, β, ξ). The log negative likelihood of GEV can be written as

$$\ell(\mu, \beta, \xi) = n\log\beta + \left(1 + \frac{1}{\xi}\right) \sum_{i=1}^{n} \log\left[1 + \xi\left(\frac{x_i - \mu}{\beta}\right)\right] + \sum_{i=1}^{n} \left[1 + \xi\left(\frac{x_i - \mu}{\beta}\right)^{\frac{-1}{\xi}}\right],$$

provided that $\left[1 + \xi\left(\frac{x_i - \mu}{\beta}\right)\right] > 0 \quad \text{for } i = 1, \ldots, n.$

2.1.2 Peaks-Over-Threshold

The block maxima method has the major defect that it is very wasteful of data. To perform an analyses only the maximum losses in large blocks are retained. For this reason it has been largely

superseded in practice by methods based on threshold exceedances, where all data that are extreme in the sense that they exceed a particular designated high level are used. Therefore, Peaks-Over-Threshold (POT) models are generally considered to be the more useful for practical applications, due to their more efficient use of the (often limited) data on extreme values in modeling the behavior of extreme values above a high threshold. An additional advantage of POT is that it provides with risk estimates that are easy to compute. Within the POT class of models one may distinguish two styles of analysis. These are the semi-parametric models built around the Hill estimator and the fully parametric models based on the generalized Pareto distribution (GPD). This paper highly concentrate more on the latter style of analysis for a reasons of relative simplicity in giving statistical estimates error using the techniques of maximum likelihood inference.

The excess distribution above the threshold u can be defined as the conditional probability

$$F_u(y) = P(X - u \leq y | X > u) = \frac{F(y+u) - F(u)}{1 - F(u)} = \frac{F(x) - F(u)}{1 - F(u)}, \qquad y > 0. \tag{2.5}$$

The methodology is based on the asymptotic approximation to the GPD of $Y = X - u$, the rescaled excesses above a suitably high level u, should the non-degenerate limiting distribution exist Pickands (1975). For those distributions F that satisfy that the distribution in (2.3) converges to (2.4), it can be shown that for large enough u there exists a positive function σ, such that (2.5) is well approximated by the the cumulative distribution function of the GPD that takes the form

$$G(y/\sigma, \xi) = \begin{cases} 1 - \left(1 + \frac{\xi}{\sigma} y\right)_+^{-\frac{1}{\xi}} & \text{if } \xi \neq 0 \\ 1 - \exp(-\frac{1}{\sigma} y) & \text{if } \xi = 0. \end{cases} \tag{2.6}$$

where $x_+ = \max(x, 0)$, defined on $\{y : y > 0\}$ and $(1 + \xi y/\sigma > 0)$ with shape parameter $\xi \in \mathbb{R}$ - known as the Extreme Value Index (EVI) and threshold dependent scale parameter σ.

The GPD is applicable to very wide classes of underlying distributions of Y according to Smith (2009). The three cases $\xi < 0$, $\xi = 0$ and $\xi > 0$ correspond to different types of tail behavior. The case $\xi < 0$ arises in distributions where there is finite upper bound on the claims which are possible in generally speaking and it might be thought that this case would apply in practice. It would be expected to detect such a limit only if there is a tendency for claims to cluster near the upper limit. The second case, $\xi = 0$, which can be obtained by a formal limit $\xi \to 0$ in (2.6) typically arises in cases with an exponentially decreasing tail. This arises not only when the distribution of Y is indeed exponential, but also from many other common distributions such as gamma, Weibull, normal, lognormal etc, so it might be expected to find the estimated value of ξ close to 0 in practice. However the third case, $\xi > 0$, which is usually referred as Pareto tail is of more concern because this corresponds to a genuinely long tailed distribution and most relevant for insurance risk managers.

In the use of GPD, a critical issue in practice is the selection of an appropriate threshold u, or equivalently the selection of an adequate number of upper order statistics. There is a trade off between bias and variance in threshold selection Bawa et al. (2001). If this is set too high so that the asymptotic theorem can be considered to be essentially exact, there will not be enough data over the threshold to calculate good estimates of σ and ξ. However, it is not actually wanted u to

be too low since there will not be point in basing the estimates on claims which are too small to be considered large claims, and to do so could induce a bias associated with lack of fit of the GPD.

There are numerous ways of choosing the threshold as well as quantifying the uncertainty of u. A diagnostic graphical tool which has been introduced by Davison and Smith (1990) is the *sample mean excess* plot, which is also known as the *sample mean residual life (MRL)* plot is a very helpful for the selection of the threshold u through visualizing description of the GPD behavior for different values of u. This is based on the fact that the mean of a GPD distributed variable Y is given by

$$E(Y) = \frac{\sigma}{1-\xi} \tag{2.7}$$

and for the introduced an excess u

$$E(Y-u|Y>u) = \frac{\sigma}{1-\xi} \tag{2.8}$$

The mean of a GPD should theoretically have linear property which means by introducing a high threshold $z > u$ should yield

$$E(Y-z|Y>z) = \frac{\sigma+\xi z}{1-\xi}, \quad \sigma+\xi z > 0 \tag{2.9}$$

which gives the average of the excesses of Y over varying values of a threshold z and is a linear transformation of (2.7). Thus the excess distribution over higher thresholds remains a GPD with the same ξ parameter but a scaling that grows linearly with the threshold z. Provided that $\xi < 1$, the mean excess function is given by

$$E(z) = \frac{\sigma+\xi(z-u)}{1-\xi} = \frac{\xi z}{1-\xi} + \frac{\sigma-\xi u}{1-\xi}, \tag{2.10}$$

where $u \leq z < \infty$ if $0 \leq \xi < 1$ and $u \leq z \leq u - \sigma/\xi$ if $\xi < 0$. The linearity of the mean excess function (2.10) in z is commonly used as a diagnostic for data admitting a GPD model for the excess distribution. It forms the basis for the following simple graphical method for choosing an appropriate threshold.

Empirically the mean excess function is defined by the points $\big(u, e_n(u)\big)$, where $e_n(u)$ is the sample mean excess function estimated as

$$e_n(u) = \frac{\sum_{i=1}^{n} (Y_i - z)\, 1_{[Y_i>u]}}{\sum_{i=1}^{n} 1_{[Y_i>u]}}. \tag{2.11}$$

and is the sum of the excesses $(Y_1 - z), \ldots, (Y_n - z)$ over the threshold z divided by the number of data points which exceeds the threshold z. The sample mean excess function describes the expected overshoot of a threshold given that exceedance occurs and is an empirical estimate of the mean excess function that is defined in (2.8). The estimated mean excess function defined in (2.9) should be linear. Whenever the points show an upward trend, it is a sign of heavy tailed behavior. Exponentially distributed data approximately would give an horizontal line and data from a short tailed distribution would show a downward trend, as noted in Corradin (2002).

Another graphical tool used to choose the threshold is the Hill graph. Let $Y_{(1)} \geq Y_{(2)} \geq \cdots \geq Y_{(n)}$ be associated descending ordered statistics of $(Y_1, Y_2, \ldots, Y_n)$ which are independent and identically distributed (iid) random variables Brilhante et al. (2013). Assuming that the distribution of these random variables is heavy-tailed, the Hill estimator Hill (1975) of tail index ξ using $k+1$ ordered statistics is defined by

$$H_{k,n} = \frac{1}{k} \sum_{i=1}^{k} \log \left(\frac{Y_{(i)}}{Y_{(k+1)}} \right). \tag{2.12}$$

Obviously, the Hill estimator is function of these extreme random variables $\{Y_{(1)}, Y_{(2)}, \ldots, Y_{(k)}\}$ which depends on the chosen threshold. A Hill plot is therefore constructed by the Hill estimator of a range of k value versus the value of k or the threshold, i.e. is defined by a set of points $\{(k, H_{k,n}^{-1}), 1 \leq k \leq n-1\}$. The value of Y_k above which the Hill estimator tends to be stable can be chosen as the optimal threshold u. The Hill estimator is closely related to the mean excess function. It is asymptotically equal to the reciprocal of the empirical mean excess function of $\log Y$ evaluated at the threshold $\log Y^{(+1)}$. An important feature of the Hill estimator to keep in mind is the variance-bias trade off that occurs when choosing the number of upper order statistics to use. Choosing too many of the largest order statistics can lead to a biased estimator, while too few increases the variability of the estimator.

2.2 Risk Measures

2.2.1 Value at Risk

Value at Risk (VaR) helps to quantify the amount of capital needed for covering loss in portfolio. It is defined as the α-th quantile of the negative returns or losses of a portfolio distribution. In other words, for some given confidence level $\alpha \in (0, 1)$. The VaR of our portfolio at the confidence level α is given by the smallest number l such that the probability that the loss L exceeds l is no larger than $(1-\alpha)$. Formally,

$$\text{VaR}_\alpha = \inf \left\{ l \in \mathbb{R} : P(L > l) \leq 1 - \alpha \right\} = \inf \left\{ l \in \mathbb{R} : F_L(l) \geq \alpha \right\}, \tag{2.13}$$

where F_L is defined as the distribution function. Typical values for α are $\alpha = 0.95$ or $\alpha = 0.99$. Note that by its definition the VaR at confidence level α does not give any information about the severity of losses which occur with a probability less than $1 - \alpha$. This is clearly a drawback of VaR as a risk measure.

2.2.2 Expected Shortfall

Expected Shortfall (ES) or the tail conditional expectation quantifies the average loss given that we have lost at least VaR. ES is computed by taking the average of losses that are larger than VaR. In other words, For a loss L with $E(|L|) < \infty$ and distribution function F_L the expected shortfall at confidence level $\alpha \in (0, 1)$ is defined as

$$\mathrm{ES}_\alpha = E\left(X|X > \mathrm{VaR}_\alpha\right) = \frac{1}{1-\alpha}\int_\alpha^1 q_u(F_L)du,$$

where $q_u(F_L) = F_L^{-1}(u)$ is the quantile function of F_L. Expected shortfall is thus related to VaR by

$$\mathrm{ES}_\alpha = \frac{1}{1-\alpha}\int_\alpha^1 \mathrm{VaR}_u(L)du$$

Instead of fixing a particular confidence level α, we average VaR over all levels $u \geq \alpha$ and thus we look further into the tail of the loss distribution. Obviously ES_α depends only on the distribution of L and obviously $ES_\alpha \geq VaR_\alpha$.

3 Preliminary Data Analysis

3.1 The data

The data used for this analysis were provided from a large database of the Ethiopian Insurance Corporation, one of the biggest insurance companies in Ethiopia. It consists of policy and claim information of vehicle insurance at the individual level. The dataset originally contains $n = 288,763$ unique individual contracts, represented by the observations $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ of $p = 10$ predictors of $\mathbf{X} = (X_1, \ldots, X_p) \in \mathbb{R}^p$ and the response variable $Y \in \mathbb{R}$ represents claim size, from July 2011 to June 2018.

3.2 Exploratory Plots for the Distribution of claim paid

The purpose of statistical graphics is to provide visual representations of quantitative and qualitative information. As a methodological tool, statistical graphics comprise a set of strategies and techniques that provide the researchers with important insights about the data under examination and help guide for the subsequent steps of the research process. The objectives of graphical methods are to explore and summarize the contents of large and complicated data sets, address questions about the variables in an analysis (for example, the distributional shapes, ranges, typical values and unusual observations), reveal structure and pattern in the data, check assumptions in statistical models, and facilitate greater interaction between the researcher and the data. Various graphical methods were examined to visualize data in raw and amalgamated formats. Additionally, beyond graphical exploratory data analysis, some methods to quantify fits of the data with some distributions are discussed.

Graphical visualization in this analysis starts with a distribution of large claims. Accordingly, more than 70% of the sum of all claims is created by only the 10% highest claims as shown in Figure 2.

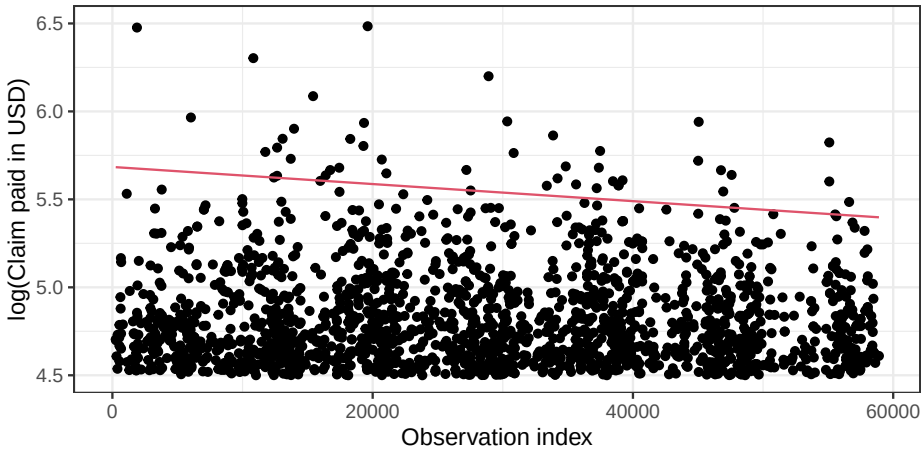

Figure 2: All claim sizes falling in an interval of $Y \in (4.5, 6.5)$ in (black points) and the 97.5% quantile (red line) of all claims.

Even for $Y \in (4.5, 6.5)$, extreme points can be seen, which indicates the important contribution of single large claims to a total risk exposure.

In Figure 2, it can be easily seen that some observations are much larger than the rest of the sample, specifically, two outliers (6.47 and 6.48) can be seen which may of course need some concern for any reason. These two largest claims could possibly happen due to a total loss, i.e. when a complete facility was lost and it may represent total losses. The second largest three claims are 6.3, 6.2 and 6.08. This shows that a particular concern should be drawn prior to the use of extreme value methods, which are applied in the rest of this section, when there is a possibility that these represent some separate process. Ideally one would like to do a separate analysis of claims resulting from total losses, but with only few such claims available, this is not practicable. The most of outliers will therefore be combined with the rest of the data for most of the analysis, but their separate origins do need to be bear in mind in interpreting the results.

The most widely recognized graphical tool to display and examine the frequency distribution and a density of a single continuous variable is the histogram. A histogram is nonparametric procedure, in a sense, constructed without assuming a statistical model and estimating its parameters from data.

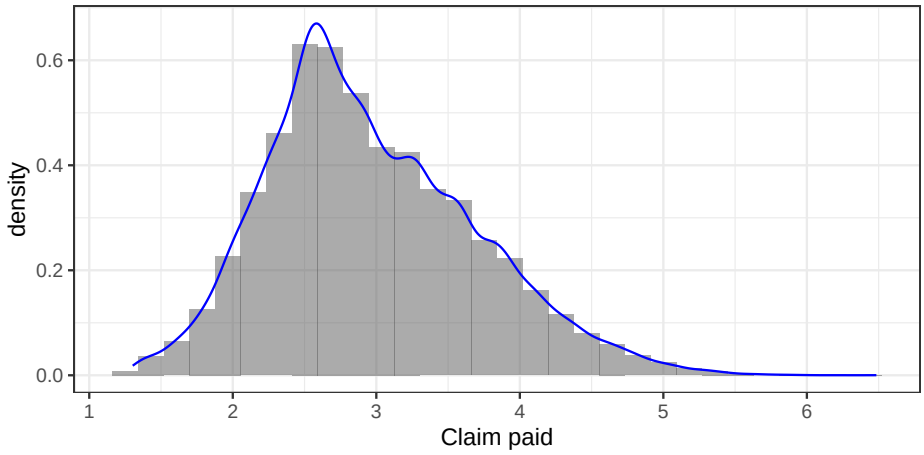

Figure 3: Frequency histogram and superimposed density plot representations of claim paid distribution.

The histogram plot in Figure 3 has a peak at the smaller values, more variability in the higher values and a bit long right tail. knowing this ahead of time is helpful to design a best fitting model later on.

Another common tool to visualize the observed distribution of data is by plotting a smoothed histogram commonly referred as empirical density, the curve superimposed on the histogram with blue line. The empirical densities overcome some of the disadvantages caused by the arbitrary discrete bins used in the basic histograms.

Although it is helpful to examine the observed data distribution, often we are examining the distribution to see whether it meets the assumption of the statistical analysis we hope to apply. As it can be seen from Figure 3, the empirical density plots shows that the claim paid variable is horizontal line (in fact some degree of smoothness is applied by default), which indicates this variable has a heavy right tail.

By proceeding claim paid examination, a quantile-quantile (Q-Q) plot can be considered as diagnostic tool to assess whether data fit or are close to a specific expected distribution. Q-Q plots can be used to judge whether observations follow a variety of distributions such as: normal, exponential and generalized Pareto distributions.

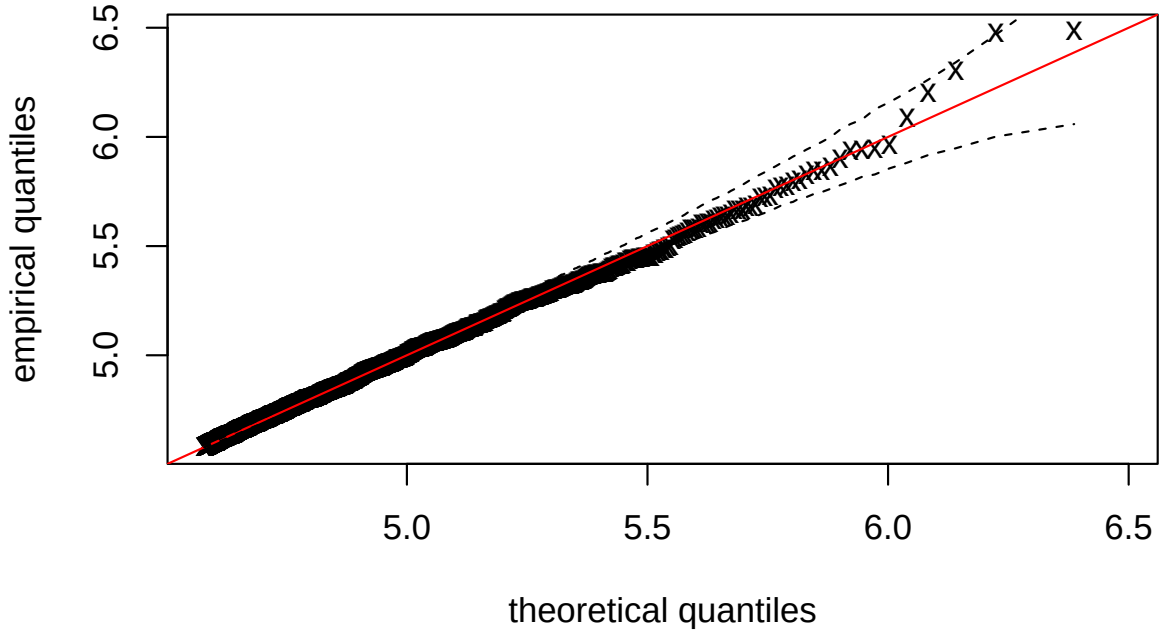

Figure 4: A Q-Q plot of claim paid empirical distribution against theoretical Generalized Pareto distribution (GPD). A threshold of $u = 4.58$ USD, above which 97.5% of claim payments fall is selected. Thus, 1,478 exceedance observations are left and the estimated parameters of GPD are: $\xi = 0.11$ and $\sigma = 0.34$

A Q-Q plot graphs the observed data quantiles against theoretical quantiles from the expected distribution. With a Q-Q plot, if the data perfectly match the expected distribution, the points will fall on a straight line. In Figure 4, we can see that the data are reasonably Generalized Pareto distributed (GPD), as all points fall fairly closely to the straight line except for few points, which can be considered as outliers. Moreover, since the plot has a curved pattern with the slope increasing from left to right, then the data has a long right tail Keen (2010), even beyond the GPD can accommodate.

Although testing whether data are consistent with a specific distribution, in this case GPD, is common, real data may be closer to many other distributions.

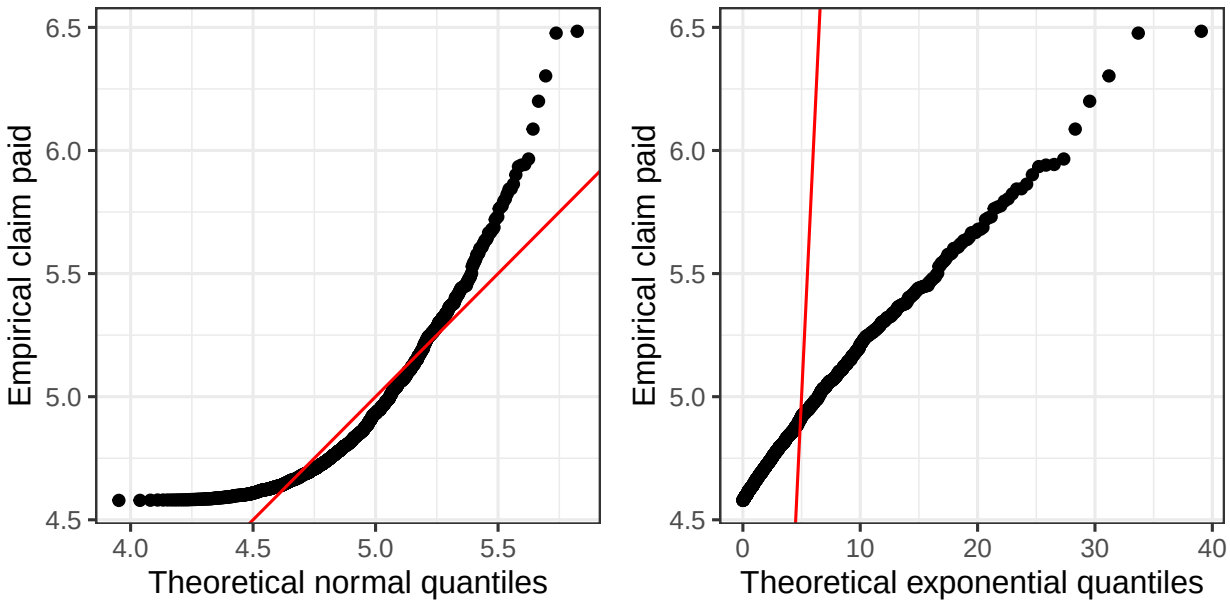

Figure 5: A normal distribution (left panel) and an exponential distribution (right panel) Q-Q plot of claim paid. The two distributions were fitted to the 1,478 exceedance observations.

The maximum likelihood estimator(s) of the parameters for normal and exponential distributions were directly computed from the empirical claim paid data.

Considering a nature of claim paid variable such as its range, a Q-Q plots shown in Figure 5 are done to evaluates the fit of claim paid variable with a specified expected quantile functions from normal and exponential distributions. The plots show that the exponential theoretical distribution for claim paid is unreliable. But in the case of normal distribution, it is some how close to that of the GPD Q-Q plot as the points are seem to be symmetric with a line.

Another way to examine whether the observed distribution appears consistent with an expected distribution is to plot the empirical density against the density for the expected distribution.

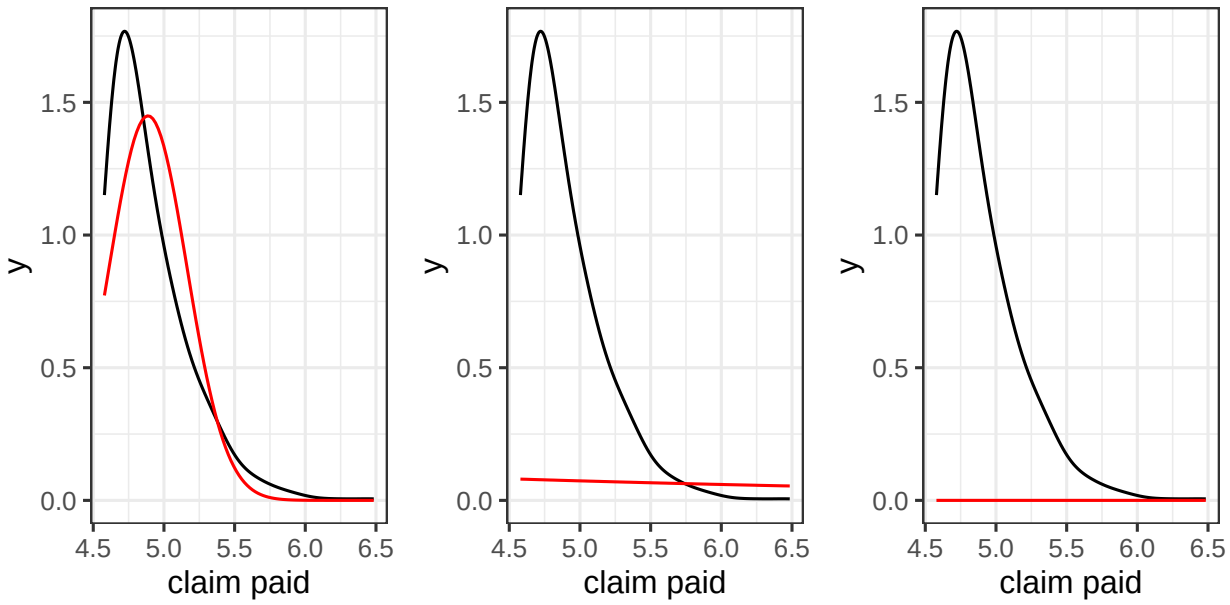

Figure 6: A theoretical curve (red line) and claim paid empirical density plot (black line) with smoothing factor of 2, from GPD (left panel), normal(middle) and exponential(right panel) distributions.

From Figure 6, it can be seen that the claim paid data appear to be close to a GPD distribution, although not perfect. For a better comparison between the normal, exponential and GPD fits, a log likelihood (LL) is employed. LL commonly used for model comparison and tells us about how likely the data are to come from that distribution with those parameters. In comparing fit of the distributions, the one that provides the higher log likelihood is a better fit for the data. Accordingly, the LL is higher for the GPD ($LL = 268.39$) than the normal distribution ($LL = -190.57$) and the exponential distribution (($LL = -3823.17$)) with only one unit difference in the degrees of freedom. These results suggest that the GPD should be picked for claim paid data. More details about the GPD model and extreme value analysis are discussed in Section 2.1.

4 Results

4.1 Modeling using GEV

The maximum claim paid data across the levels of manufacturer company are shown in Figure 7. No obvious trend is observed.

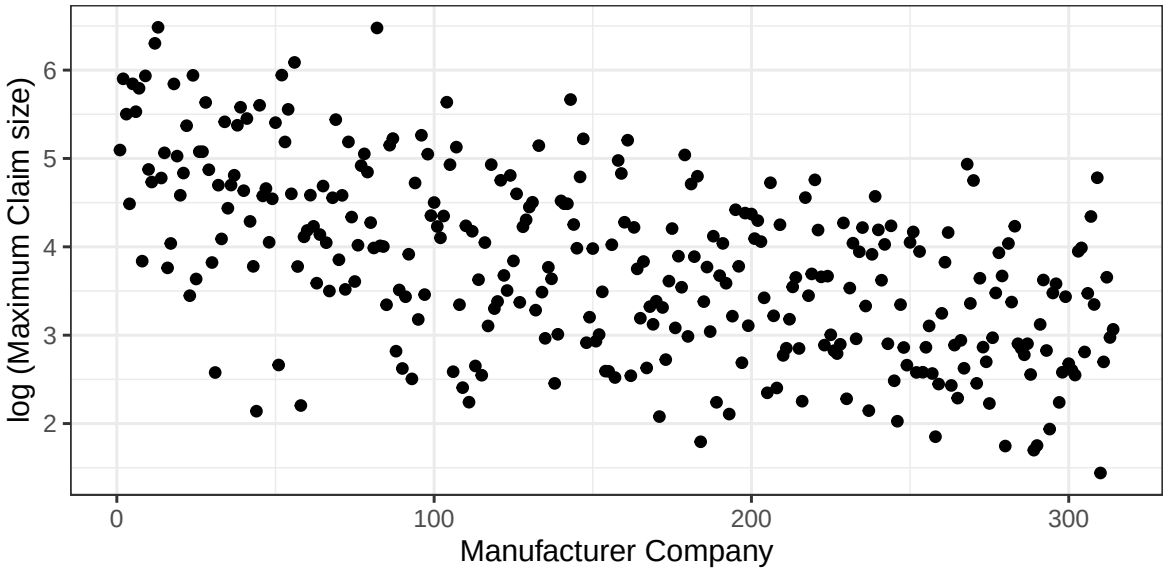

Figure 7: Scatter plot of maxima claim paid across manufacture company for Ethiopian vehicle insurance dataset.

Figure 7 represents maximum claim paid across manufacturer company, showing extreme claim paid using block maxima method. The highest four amount of claim paid, which seem to be outliers occurred in USD are 3.05m, 2.99m, 2.01m and 1.22m, that correspond to FIAT, SKY BUS, IVECO and BISHOFTU, respectively.

For GEV estimation, the Block Maxima of maximum claim paid across manufacturer company are extracted. The blocks $n = 314$ types of manufacturer company have been chosen to be reasonably large, so the GEV model is fitted to the $n = 314$ across the manufacturer's maxima using maximum likelihood estimation. The point MLE of the parameters $(\mu; \beta; \xi)$ for the GEV distribution and their respective 95% CI's are summarized in Table 4.1. Based on these estimates, VaR is calculated by (2.13) and VaR estimate is also presented in Table 4.1.

	Lower bound	Point Estimate	Upper bound
$\hat{\mu}$	3.30	3.42	3.54
$\hat{\beta}$	0.91	0.99	1.08
$\hat{\xi}$	0.14	0.22	0.30
$\widehat{\text{VaR}}_{0.01}$	–	12.96	–
$\widehat{\text{ES}}_{0.01}$	–	14.52	–

Table 4.1: Point and 95% CI estimates of GEV parameters, and point estimates of risk measures.

The CI results shows that the confidence interval of ξ does not contains 0 and both lower and upper bounds are positive, which means the Fréchet distribution could be a more accurate model in the entire GEV family.

Since we have our estimated parameters of GEV, we can calculate the risk measures (VaR and ES), which are contained in Table 4.1. Notice that at 99% VaR and ES are 12.96 and 14.52, which indicate that, with probability 0.01, the insurance company makes claim payment of at least 12.96 and on a long position the claim payment will reach up to 14.52 on an average.

4.2 Modeling using GPD

The mean excess plot in left panel of Figure 8 is in fact fairly linear over the entire range of the claim size distribution and its upward slope leads us to expect that a GPD with positive shape parameter ξ could be fitted to the entire dataset. However, there is some evidence of a "kink" in the plot below the value $285,000$ and a straightening out of the plot above this value, so we have chosen to set our threshold at $u = 285,000$ and fit a GPD to excess claim sizes of 58 observations above this threshold, in the hope of obtaining a model that is a good fit to the largest of the claim sizes. If the data really follow a GPD, then this plot should stay close to a straight line of slope $\xi/(1-\xi)$, provided $\xi < 1$ (Smith, 2009). The apparent exception to linearity of the mean excess plot is at the right-hand end of the plot, but in fact this is not such a significant matter because in this region there are very few data points- the mean excess is computed from a very small number of exceedances and hence has a lot of sampling variability. On the basis of this plot, the evidence in favour of the GPD seems good. The ML parameter estimates are $\hat{\xi} = 0.4$ and $\hat{\sigma} = 5.31$ with standard errors 0.15 and 3.62, respectively. Thus the model we have fitted is essentially a heavy-tailed, infinite-variance model. A picture of the fitted GPD model for the excess distribution $\hat{F}_u(y-u)$ is also given in right panel of Figure 8, superimposed on points plotted at empirical estimates of the excess probabilities for each claim size; note the good correspondence between the empirical estimates and the GPD curve. The Hill estimator (the reciprocal of ξ) in Figure 9 concise with the mean

excess plot, indicating that the Hill estimator starts to be stable after the vertical red line, which is drawn at threshold of $u = 285,000$. The estimates of tail index α obtained are between 1.25 and 2.5, suggesting ξ estimates between 0.4 and 0.8, all of which correspond to infinite-variance models for these data. The tail index α estimates based on $k = 50, \ldots, 120$ order statistics mostly range from 1.25 to 2.5, suggesting a ξ value in the range $0.4 - 0.8$, which is larger than the values estimated in with a GPD model.

In Figure 9, it can be noted that the high variability in the left region (the one determined by the largest order statistics) of the plot is not a welcome feature, since it makes difficult the proper selection of the number of upper order statistics involved in the estimation of the tail index. An important question that often arises in practice is whether one should ignore those observations, thus ignoring useful information about the behavior of the tail, or include them and get a biased estimate of α. Even though the values of α seem to be decrease as a number of exceedences increase, it can be seen that for the ideal case of setting, to a large extent the plots perform satisfactorily allowing the data analyst to identify correctly the underlying value of the tail index.

In insurance we might use the model to estimate the expected size of the insurance claim, given that it enters a given insurance layer. Thus we can estimate the expected claim size given exceedance of the threshold of USD 285, 000 or of any other higher threshold by using (2.10) with the appropriate parameter estimates.

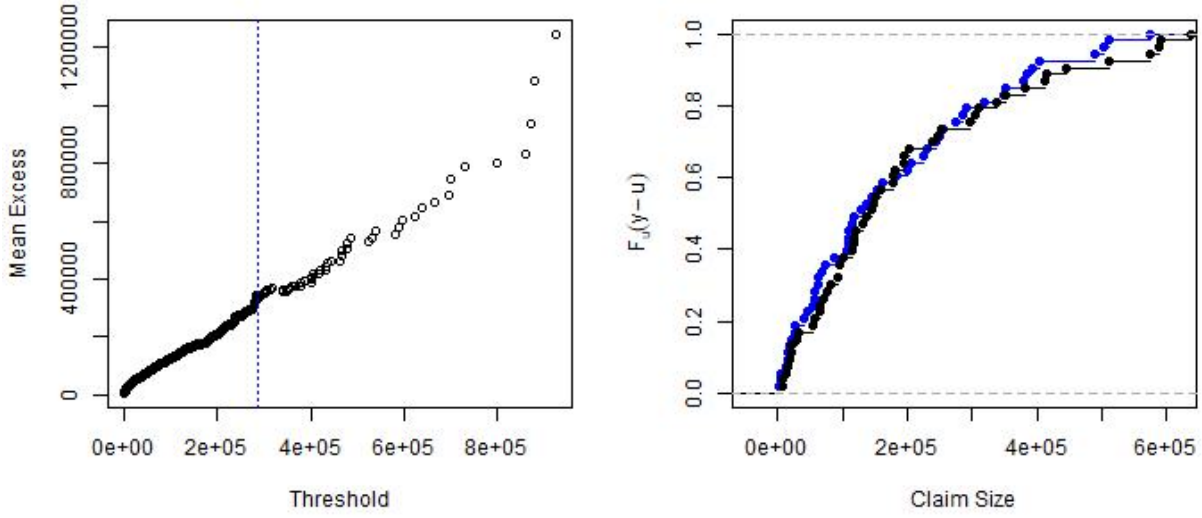

Figure 8: Sample mean excess plot (left panel) and on the right panel is Empirical distribution of excesses(black points) and fitted GPD (blue points).

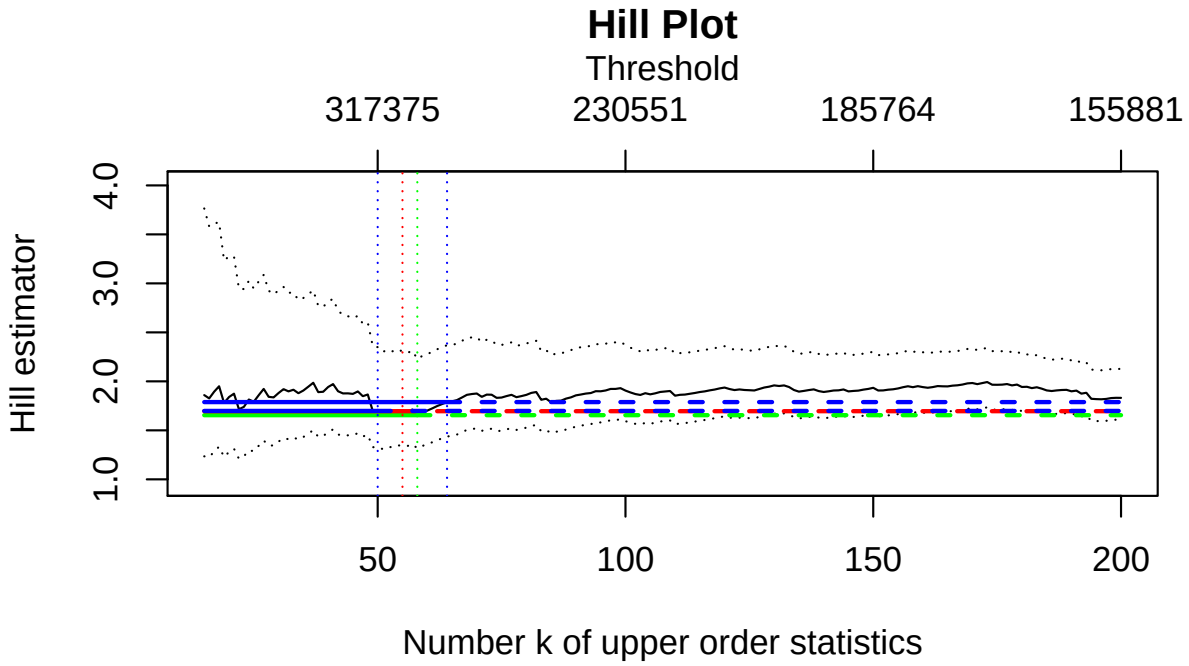

Figure 9: Hill plots of four different thresholds together with 95% confidence interval, and their respective estimated the tail indices $\alpha = 1/\xi's$. The four ξ estimators obtained are $\hat{\xi} = 0.56, 0.6, 0.59, 0.57$ by setting the corresponding thresholds $u = 280000, 285000, 300000, 315000$. The number of exceedences k is plotted on the horizontal axis while the estimation of the reciprocal of ξ is plotted on vertical axis.

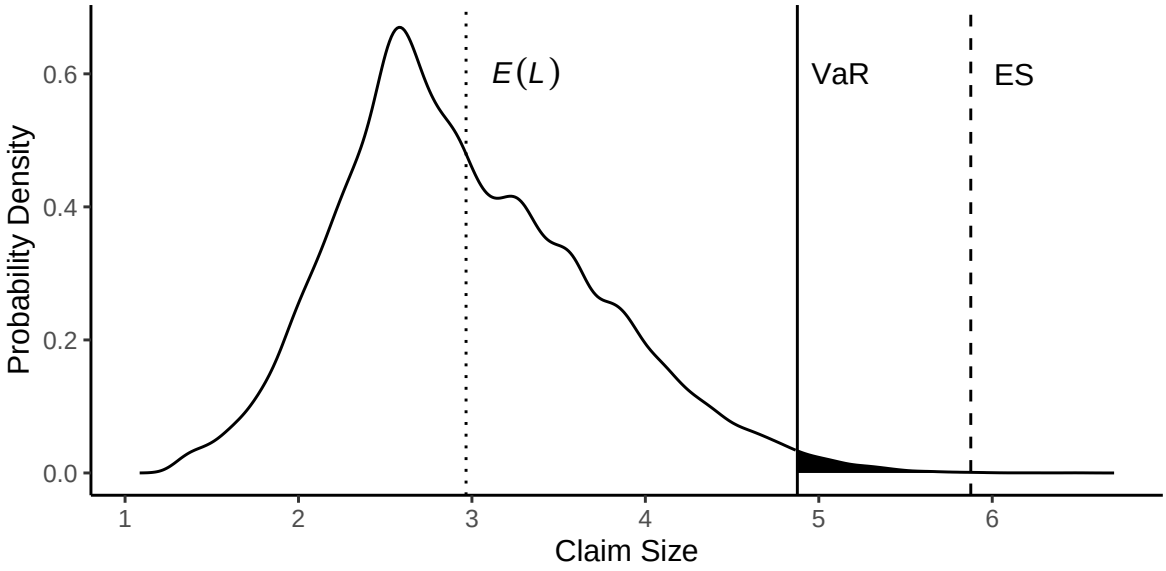

Figure 10: Claim size distribution with the 99% VaR marked as a vertical line; the mean claim is shown with a dotted line and the 99% expected shortfall is marked with a dashed line.

Figure 10 illustrates the notion of VaR. The probability density function of a claim size distribution is shown with a vertical line at the value of the 99% VaR. Note that the estimated mean claim size is ($E(L) = 3.7$), while the 99% estimated VaR and ES values are approximately 5.84 and 5.87, respectively; indicating that there is a 1% chance that the insurance company makes claim payment of at least 5.84. Denoting by μ the mean of the claim size distribution, sometimes the statistic $\text{VaR}_\alpha^{\text{mean}} := \text{VaR}_\alpha - \mu$ is used for capital-adequacy purposes instead of ordinary VaR. The distinction between ordinary VaR and $\text{VaR}_\alpha^{\text{mean}}$ is of little relevance in market risk management, where the time horizon is short and μ is close to zero. It becomes relevant in credit where the risk-management horizon is longer. In particular, in loan pricing one uses VaR^{mean} to determine the economic capital needed as a buffer against unexpected losses in a loan portfolio. Taking the expectation of the claim size distribution into account is also important in the growing field of asset-management risk. The 99% expected shortfall value is 5.87, which is much higher than the expected claim size value of 3.7 in this case.

5 Discussion and Conclusion

The purpose of this study was to analyze the tail distribution of claim size for Ethiopian vehicle insurance dataset using EVT. We used both GEV and GPD distributions in block maxima and peaks-over-threshold methods, respectively. GEV and GPD are two different distributions, but in this context their purpose is the same; they show the distribution of extreme claim size. Also, the shape parameter ξ is the same parameter in the two distributions. Theoretically, we should get the same estimation of ξ in the methods (Coles, 2001). A weaker hypothesis is that the sign of ξ should

be the same in the models, which the case in this study. This is also the case in Gilli and Këllezi (2006). Even though we obtained $\hat{\xi} > 0$ in both models, the magnitudes are different; $\hat{\xi} = 0.22$ in GEV and $\hat{\xi} = 0.4$ in GPD.

One of the objectives of this paper is to answer the question, How much can the claim size fall beyond certain level of threshold. To answer this question, we use the two risk measures VaR and ES. However, how the risk measures should be estimated is not straightforward; there exist several methods for this purpose [see Hull (2018)]. Here, we focus on the block maxima and POT methods. Before estimating VaR and ES, the parameters of the distributions must be estimated: $(\mu; \beta; \xi)$ in GEV and $(\sigma; \xi)$ in GPD. Since we obtained different estimators of ξ in GEV and GPD, we obviously get different estimators of VaR and ES. The 99% VaR and ES estimators in GEV are 12.96 and 14.52, respectively, while in the GPD, the estimators are 5.84 and 5.87, respectively. This arises an interesting question such as; which one of the the two methods that produces the more accurate estimates of VaR and ES. Performing the "Backtesting" strategy is the popular approach to evaluate the estimates of VaR and ES, which is beyond the scope of this paper and the next step research of the author.

It is possible that the choice of α influence the result. A small α needs to be chosen for the formulas of VaR and ES to be accurate [see Dowd (2005)]. Here, we let $\alpha = 0.01$, but an even smaller α should be even better. Since POT extract the extreme events more efficient than GEV, it is possible that POT is more sensitive to the choice of α than GEV. A solution is then to chose a smaller α . On the other hand, this would imply that fewer observations will be considered as extreme, which also can lead to poor estimates. Choosing another value of α could also bring some new light on the discussion.

References

J. Bawa, L. Trenner, S. Coles, and P. Dorazio. *An introduction to statistical modeling of extreme values*. Springer, 2001.

M. F. Brilhante, M. I. Gomes, and D. Pestana. A simple generalisation of the hill estimator. *Computational Statistics & Data Analysis*, 57(1):518–535, 2013.

S. Coles. An introduction to statistical modeling of extreme values. , 2001.

S. Corradin. Economic risk capital and reinsurance: an extreme value theory's application to fire claims of an insurance company. Sixth International Congress on Insurance: Mathematics and Economics, Lisbon, 2002, 2002.

A. C. Davison and R. L. Smith. Models for exceedances over high thresholds (with discussion). *Journal of the Royal Statistical Society, Series B (Methodological)*, 52(3), 1990.

K. Dowd. Measuring market risk. , 2005.

R. Fisher and L. Tippet. Limiting forms of the frequency distribution of the largest or smallest member of a sample. 1928.

M. Gilli and E. Këllezi. An application of extreme value theory for measuring financial risk. *Computational Economics*, 27(207 - 228):2 – 3, 2006.

B. M. Hill. A simple general approach to inference about the tail of a distribution. *Annals of Statistics*, 13, 1975.

J. C. Hull. Risk management and financial institutions. , 2018.

K. J. Keen. *Graphics for Statistics and Data Analysis with R*. Texts in Statistical Science. CRC Press, Taylor & Francis group, Chapman & Hall/CRC, NY, USA, 2010.

J. McNeil, P. Embrechts, and R. Frey. Quantitative risk management concepts, techniques and tools. *Princeton University Press Princeton and Oxford*, 2005.

J. I. Pickands. Statistical inference using extreme value order statistics. *Annals of Statististics*, 1975.

R. L. Smith. *Extreme value analysis of insurance risk*. Department of Statistics and Operations Research, University of North Carolina,Chapel Hill, NC 27599-3260, 2009.

Extremal Random Forests

Nicola Gnecco*[1], Edossa Merga Terefe*[1, 2], and Sebastian Engelke[1]

[1]Research Center for Statistics, University of Geneva, Switzerland
[2]Statistics Department, Hawassa University, Ethiopia
[1] {*nicola.gnecco, edossa.terefe, sebastian.engelke*} *@unige.ch*

Abstract

Classical methods for quantile regression fail in cases where the quantile of interest is extreme and only few or no training data points exceed it. Asymptotic results from extreme value theory can be used to extrapolate beyond the range of the data, and several approaches exist that use linear regression, kernel methods or generalized additive models. Most of these methods break down if the predictor space has more than a few dimensions or if the regression function of extreme quantiles is complex. We propose a method for extreme quantile regression that combines the flexibility of random forests with the theory of extrapolation. Our extremal random forest (ERF) estimates the parameters of a generalized Pareto distribution, conditional on the predictor vector, by maximizing a local likelihood with weights extracted from a quantile random forest. Under certain assumptions, we show consistency of the estimated parameters. Furthermore, we penalize the shape parameter in this likelihood to regularize its variability in the predictor space. Simulation studies show that our ERF outperforms both classical quantile regression methods and existing regression approaches from extreme value theory. We apply our methodology to extreme quantile prediction for U.S. wage data.

Keywords: extreme quantiles; local likelihood estimation; quantile regression; random forests; threshold exceedances.

1 Introduction

Quantile regression is a well-established technique to model statistical quantities that go beyond the conditional expectation that is used for standard regression analysis (Koenker and Bassett, 1978). This is particularly valuable in applications such as economics, survival analysis, medicine, and

*Authors contributed equally.

finance (Angrist et al., 2006; Yang, 1999; Heagerty and Pepe, 1999; Taylor, 1999; Yu et al., 2003), where one needs to model the heteroschedasticity of the response or conditional quantiles such as the median.

In this paper, we consider the problem of estimating high conditional quantiles of a response variable $Y \in \mathbb{R}$ given a set of predictors $X \in \mathbb{R}^p$ in large dimensions, an important task in risk assessment for rare events (Chernozhukov, 2005). For a fixed predictor value x, define $Q_x(\tau)$ as the quantile at level $\tau \in (0, 1)$ of the conditional distribution of $Y \mid X = x$. We are interested in the estimation of extreme quantiles where $\tau \approx 1$ is close to one. This estimation problem exhibits two fundamental challenges that are illustrated in Figure 1, which shows a simulation similar to Athey et al. (2019, Figure 2). The predictor space has $p = 40$ dimensions and only the first variable X_1 has a signal corresponding to a scale shift in Y; see Example 1 in Section 3.1 for details.

The first challenge in estimating $Q_x(\tau)$ relates to the fact that for an extreme probability level, say $\tau = 0.9995$ as in Figure 1, there are typically only few or no observations in the sample that exceed the corresponding conditional τ-quantiles. Indeed, for a sample of size n, the expected number of exceedances above the conditional τ-quantile is $n(1 - \tau)$, which becomes smaller than one if $\tau > 1 - 1/n$. Therefore, using an empirical estimator based on quantile loss leads to a large bias. A second challenge stems from the possibly high-dimensional predictor space $\mathbb{R}^p$, where there might be no training observations close to x; note that the Figure 1 only shows the first of the 40 dimensions of X. Too simple regression models may then introduce additional bias.

The first challenge can be addressed by relying on tail approximations motivated by extreme value theory (e.g., de Haan and Ferreira, 2006), which allow the extrapolation to quantile levels beyond the range of the data. Such methods typically consider (transformations of) linear (Chernozhukov, 2005; Wang and Tsai, 2009; Wang et al., 2012; Wang and Li, 2013) functions, additive models (Chavez-Demoulin and Davison, 2005; Youngman, 2019), non-parametric regression (Beirlant et al., 2004; Martins-Filho et al., 2015) and local smoothing methods (Daouia et al., 2011; Gardes and Stupfler, 2019; Velthoen et al., 2019). However, these existing approaches are either not flexible enough to model complex response surfaces or do not scale well in higher dimensions p of the predictor space.

Regarding the second challenge, several quantile regression methods have been proposed in the statistical and machine learning literature that can cope with high-dimensional predictor spaces and complex regression surfaces (Taylor, 2000; Friedman, 2001). In particular, there exist several forest-based approaches for quantile regression (Meinshausen, 2006; Athey et al., 2019). These methods are based on (extensions of) the random forest originally developed by Breiman (2001) and can estimate flexible quantile regression functions. Compared to methods such as gradient boosting and neural networks, the main advantage of forest-based approaches is that they require little tuning and that their statistical properties are relatively well understood (Athey et al., 2019). Moreover, they scale well with the dimension of the predictor space as opposed to approaches based on generalized additive models (Koenker, 2011) and kernel-based methods (Yu and Jones, 1998). While these methods work well for estimation of quantiles inside the data range, such as $\tau_0 = 0.8$ in Figure 1, their performance deteriorates for quantile estimation at extreme levels $\tau \approx 1$ close to the upper endpoint of the response distribution.

In this paper, we bring together ideas from extreme value theory and forest-based regression

methods to tackle the challenges of extreme quantile regression in predictor spaces with possibly high dimensions p. To extrapolate beyond the data range, we rely on the approximation by the generalized Pareto distribution (GPD) of the exceedances over an intermediate threshold; see the triangles in Figure 1. Under mild assumptions, the conditional quantile of Y, given $X = x$, at level $\tau \approx 1$ can the be approximated by (Balkema and de Haan, 1974; Pickands, 1975)

$$Q_x(\tau) \approx Q_x(\tau_0) + \frac{\sigma(x)}{\xi(x)} \left[\left(\frac{1-\tau}{1-\tau_0} \right)^{-\xi(x)} - 1 \right], \tag{1.1}$$

where $Q_x(\tau_0)$ is an intermediate quantile at level $\tau_0 < \tau$ and the second term on the right-hand side is quantile function of the GPD indexed by the conditional scale $\sigma(x) > 0$ and shape parameter $\xi(x) \in \mathbb{R}$. This includes responses with heavy tails ($\xi(x) > 0$), light tails ($\xi(x) = 0$) and with finite upper end points ($\xi(x) < 0$). The intermediate quantile level τ_0 is chosen small enough such that the conditional quantiles $Q_x(\tau_0)$ can be estimated by classical regression methods. At the same time, it should be large enough so that the approximation in (1.1) by the GPD is accurate.

In order to cope with complex response surfaces and high-dimensional predictor spaces, we rely on ideas from the random forest literature (Meinshausen, 2006; Athey et al., 2019). Our new extremal random forest (ERF) localizes the estimation of the GPD parameter vector $\theta(x) = (\sigma(x), \xi(x))$ around the predictor value x using forest-based weights. Since only few extreme observations are typically available for training, the simple tuning of random forests turns out to be of great advantage. Under certain conditions, we show consistency of the ERF estimator $\hat{\theta}(x)$ for the true conditional parameter vector $\theta(x)$. Since our loss function, namely the GPD log-likelihood, is non-convex, the proof strategy of Athey et al. (2019) cannot be used, and we rely on the theory of Newey (1991).

Our ERF algorithm combines the advantages of accurate tail extrapolation at levels $\tau \approx 1$ with a flexible regression method that scales well with predictor dimension. In simulations, we show that ERF outperforms extreme value theory and quantile regression methods to estimate extreme quantiles. Moreover, it is competitive with the recent gradient boosting by Velthoen et al. (2021) and has the advantage of significantly easier tuning and the theoretical guarantee of our consistency result. Finally, we apply our methodology to extreme quantile prediction for U.S. wage data (Angrist et al., 2009). The ERF algorithm is available as an `R` package on `https://github.com/nicolagnecco/erf`.

2 Background

2.1 Extreme Value Theory

The first challenge of extreme quantile regression is that only a few or even no data points exceed the quantiles of interest. This section considers the classical case of unconditional extremes without predictors. Let $Y_1, \ldots, Y_n$ be n independent copies of a real-valued random variable Y. The notion of an extreme quantile $\tau = \tau_n$ is typically expressed relative to the sample size n. The expected number of observations in the sample that exceed the τ_n-quantile is then $n(1 - \tau_n)$. A quantile with

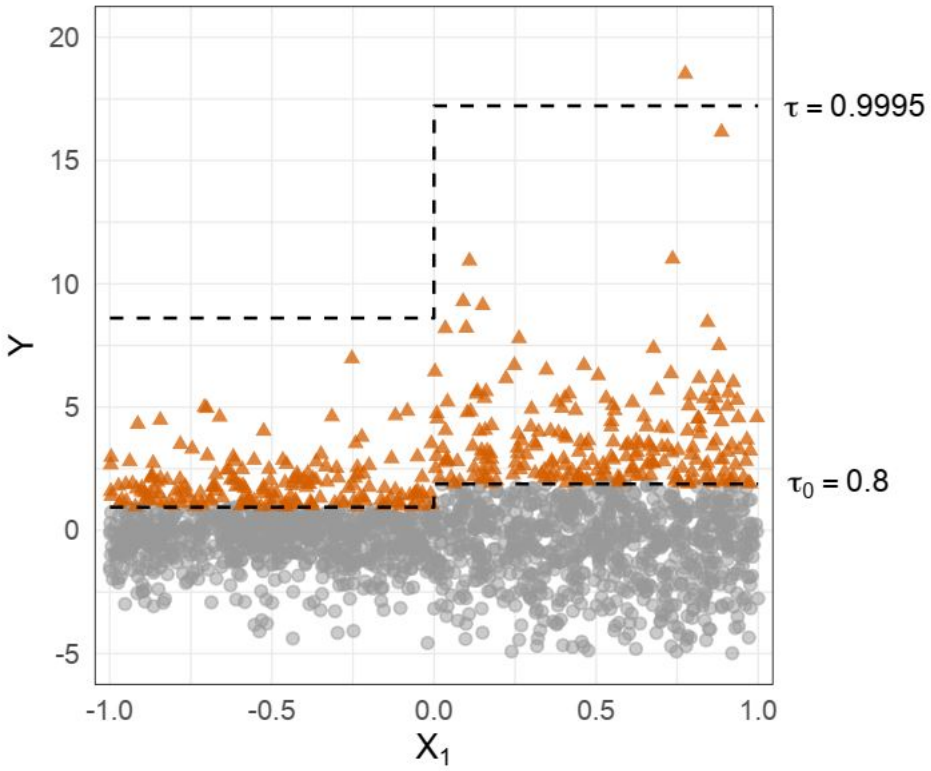

Figure 1: Realization of $n = 2000$ samples from the generative model in Example 1 in Section 3.1. Response Y is plotted against the first predictor X_1. Dashed lines represent the quantile functions associated to the intermediate $\tau_0 = 0.8$ and high $\tau = 1 - 1/n = 0.9995$ quantile levels. Triangles are observations above the intermediate threshold.

level $\tau_n \to 1$ such that $n(1 - \tau_n) \to \infty$ is called an intermediate quantile. Empirical estimation in this case still works well since the effective sample size, that is, the number of exceedances, grows to infinity (de Haan and Ferreira, 2006). For risk assessment, the most critical case is if the quantile of interest is eventually beyond the range of the data, that is, $(1 - \tau_n)n \to 0$ as $n \to \infty$. Then, we can no longer rely on empirical estimators but must resort to asymptotically motivated approximations from extreme value theory.

Let $u^* \in (0, \infty]$ be the upper endpoint of the distribution of Y. Under mild regularity assumptions on the tail of Y, the Pickands–Balkema–De Haan theorem (Balkema and de Haan, 1974; Pickands, 1975) states that there exists a normalizing function $\sigma(u) > 0$ such that

$$\lim_{u \to u^*} \mathbb{P}\left(\frac{Y - u}{\sigma(u)} \leq z \mid Y > u \right) = G(z; (1, \xi)), \tag{2.1}$$

where the limit on the right-hand side is the distribution function of the generalized Pareto distribution (GPD) (Pickands, 1975) given by

$$G(z; \theta) = 1 - \left(1 + \frac{\xi}{\sigma} z \right)_+^{-1/\xi}, \quad z > 0, \tag{2.2}$$

and $\theta = (\sigma, \xi) \in (0, \infty) \times \mathbb{R}$ is the parameter vector consisting of scale and shape, respectively. The shape parameter $\xi \in \mathbb{R}$, also known as the extreme value index (Beirlant et al., 2005), characterizes the decay of the tail of Y. If $\xi > 0$, then Y is heavy-tailed; if $\xi = 0$, then Y is light-tailed; if $\xi < 0$ then Y has a finite upper endpoint. Moreover, the GPD is a natural model for the distribution tails since it is the only possible limit of threshold exceedances as in (2.1).

The GPD approximation can be directly translated into an approximation for the small probability of Y exceeding a high threshold y. By Bayes' theorem and (2.1) we obtain

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

where $u < y$ denotes an intermediate threshold. Combining this approximation with (2.2) and letting $\mathbb{P}(Y > y) = 1 - \tau$ and $\mathbb{P}(Y > u) = 1 - \tau_0$, we obtain an approximation for the τ-quantile of Y as

$$Q(\tau) \approx Q(\tau_0) + \frac{\sigma}{\xi}\left[\left(\frac{1-\tau}{1-\tau_0}\right)^{-\xi} - 1\right], \tag{2.3}$$

where $Q(\tau_0) := F_Y^{-1}(\tau_0)$ denotes the intermediate quantile at level $\tau_0 < \tau$.

In applications, the scale and shape parameters of the GPD have to be estimated from independent observations $Y_1, \ldots, Y_n$ of Y. We fix an intermediate quantile level τ_0 and define the exceedances $Z_i = (Y_i - \hat{Q}(\tau_0))_+, i = 1, \ldots, n$, where $\hat{Q}(\tau_0)$ denotes the empirical τ_0 quantile. We can estimate the GPD parameter vector θ by maximum-likelihood, where the negative log-likelihood (or deviance) contribution of the ith exceedance Z_i is

$$\ell_\theta(Z_i) = \log\sigma + \left(1 + \frac{1}{\xi}\right)\log\left(1 + \frac{\xi}{\sigma}Z_i\right), \quad \theta \in (0, \infty) \times \mathbb{R}, \tag{2.4}$$

if $Z_i > 0$, and zero otherwise.

2.2 Quantile Regression and Generalized Random Forests

Given a pair (X, Y) of predictor vector $X \in \mathbb{R}^p$ and response variable $Y \in \mathbb{R}$, quantile regression deals with modeling the conditional τ-quantile $Q_x(\tau)$ of the conditional distribution of Y given that $X = x$ for a particular predictor value $x \in \mathbb{R}^p$. The main challenge is that the dimension p of the predictor space may be large and that the quantile surface $Q_x(\tau)$ as a function x may be a complex, highly non-linear function.

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be n independent copies of the random vector (X, Y). In contrast to the setting in Section 2.1, classical methods for quantile regression consider a fixed quantile level $\tau \equiv \tau_n$ that does not change with the sample size. On population level, these methods exploit the fact that the conditional quantile function is the minimizer of the expectation of the quantile loss $\rho_\tau(c) = c(\tau - \mathbb{1}\{c < 0\})$, $c \in \mathbb{R}$, (Koenker and Bassett, 1978), that is,

$$Q_x(\tau) = \arg\min_{q \in \mathbb{R}} \mathbb{E}[\rho_\tau(Y - q) \mid X = x]. \tag{2.5}$$

The above expectation cannot be estimated directly on the sample level since the set of observed predictor values does not typically include the value x. A natural estimator is

$$\hat{Q}_x(\tau) = \arg\min_{q \in \mathbb{R}} \sum_{i=1}^{n} w_n(x, X_i)\rho_\tau(Y_i - q), \tag{2.6}$$

where $x' \mapsto w_n(x, x')$ is a set of localizing similarity weights around the predictor value of interest. The weights can for instance be obtained by a kernel approach (Yu and Jones, 1998), but this is limited to moderately large dimensions (Stone, 1980, 1982).

In order to model more complex quantile surfaces in larger dimensions, Meinshausen (2006) and Athey et al. (2019) propose to use the estimator (2.6) with similarity weights $w_n(\cdot, \cdot)$ obtained from a random forest. Random forests (Breiman, 2001) are an ensemble method used for both regression and classification tasks and consist of fitting B decision trees to the training data. In regression settings, each decision tree predicts a test point $x \in \mathbb{R}^p$ by

$$\mu_b(x) := \sum_{i=1}^{n} \frac{\mathbb{1}\{X_i \in L_b(x)\}Y_i}{|\{i : X_i \in L_b(x)\}|}, \quad b = 1, \ldots, B,$$

where $L_b(x) \subset \mathbb{R}^p$ denotes the rectangular region that x belongs to in bth tree. By defining the similarity weights $w_{n,b}(x, X_i) := \mathbb{1}\{X_i \in L_b(x)\}/|\{i : X_i \in L_b(x)\}|$, the random forest predictions can be written as

$$\mu(x) := \frac{1}{B}\sum_{b=1}^{B} \mu_b(x) = \sum_{i=1}^{n} w_n(x, X_i)Y_i,$$

where $w_{n,b}(x, X_i) = \sum_{b=1}^{B} w_{n,b}(x, X_i)/B$ is the average weight across B trees.

The original idea of Meinshausen (2006) is to use the weights estimated by this standard regression random forest for quantile regression in (2.6). A drawback of this approach is that the similarity weights arise from decision trees that are grown by minimizing the mean squared error loss. This leads to the fact that, as stated in Meinshausen (2006), $w_n(x, X_i)$ takes large values for those observations i such that $\mathbb{E}[Y \mid X = X_i] \approx \mathbb{E}[Y \mid X = x]$. In many situations the conditional expectation is not representative of the whole conditional distribution of $Y \mid X = x$, and it may happen that $w_n(x, X_i)$ is large but $Q_{X_i}(\tau) \not\approx Q_x(\tau)$; see Athey et al. (2019, Figure 2) or our Figure 1 where the conditional expectation is constant over the predictor space. In these cases, the similarity weights estimated with standard random forest do not capture the heterogeneity of the quantile function and are thus not well-suited for quantile regression tasks. Athey et al. (2019) introduced generalized random forests (GRF), a method designed to fit random forests with custom loss functions. The GRF retains all the appealing features of classical random forests, i.e., it is simple to fit and requires little tuning of hyperparameters. One of the main applications of GRF is quantile regression, where the trees of the forest are grown to minimize the quantile loss function. In this work, we rely on GRF with quantile loss to estimate similarity weights $w_n(\cdot, \cdot)$ that capture the variation of the entire conditional distribution of $Y \mid X = x$ in the predictor space. In practice, the GRF algorithm estimates simultaneously conditional quantiles at levels $\tau = 0.1, 0.5, 0.9$ as a proxy for the conditional distribution of $Y \mid X = x$. For simplicity, in the sequel, we refer to GRF with quantile loss as GRF.

3 Extremal Random Forest

3.1 The Algorithm

In this work we study a method for flexible extreme quantile regression where both challenges described in Sections 2.1 and 2.2 occur simulateneously. Consider the random vector (X, Y) of predictors $X \in \mathcal{X} \subset \mathbb{R}^p$ and response $Y \in \mathbb{R}$, with $\mathcal{X}$ compact. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be independent copies of (X, Y). In many applications in risk assessment, the goal is to estimate the quantile function $x \mapsto Q_x(\tau) = F^{-1}_{Y|X=x}(\tau)$, at an extreme level $\tau = \tau_n$, where the expected number of observations in the sample that exceed their conditional quantiles is small and possibly tends to 0 as $n \to \infty$; see Section 2.1. To illustrate the challenges of this estimation problem, we consider an example where the scale of the response variable Y is modeled as a step function of the covariates X. This corresponds to Athey et al. (2019, Figure 2), except that we assume that the noise of the response variable is heavy-tailed instead of Gaussian.

Example 1. Let $X \sim U_p$ be a uniform distribution on the cube $[-1, 1]^p$ in dimension p and $Y \mid X = x \sim s(x)\, T_4$, where T_ν denotes a Student's t-distribution with $\nu > 0$ degrees of freedom. The shape parameter of the conditional distribution $Y \mid X = x$ is then constant $\xi(x) = 1/\nu(x) \equiv 0.25$ and we choose the $s(x) = 1 + \mathbb{1}\{x_1 > 0\}$ for $x \in \mathbb{R}^p$. The GPD scale parameter $\sigma(x)$ of $Y \mid X = x$ and therefore also the quantile function $Q_x(\tau)$ only depend on X_1. The other predictors are noise variables.

Figure 1 in the introduction shows $n = 2000$ observations sampled from the model of Example 1 in dimension $p = 40$. The goal is to predict the conditional quantile $Q_x(\tau)$ for a high level of τ, e.g., $\tau = 0.9995$. We observe that the difficulty of the task is twofold. First, because of a possibly high-dimensional predictor space, there might be no training observations close to x; note that we only show the first of the 40 dimensions of X in the figure. Second, the τ-quantile might be out of the range of the data if τ is very close to one. Indeed, for a sample of size n, the expected number of exceedances above the conditional τ-quantile is $n(1 - \tau)$, which becomes smaller than one if $\tau > 1 - 1/n$.

Our methodology accurately addresses both of these challenges. For effective localizing in the predictor space, even in high-dimensional problems, we use the weights emerging from GRF (Athey et al., 2019). For correct extrapolation in the tail of the conditional response variable, we rely on the asymptotic theory of extremes and fit a localized generalized Pareto distribution; see Section 2.1. More precisely, for an intermediate quantile level τ_0, we assume that the distribution function of $Y - Q_x(\tau_0)$, conditional on $Y > Q_x(\tau_0)$, is approximately generalized Pareto (Balkema and de Haan, 1974) with scale and shape parameters depending on the predictor value x, that is, for any $z > 0$,

$$\mathbb{P}\left(Y - Q_x(\tau_0) \le z \mid Y > Q_x(\tau_0), X = x\right) \approx G\left(z; \theta(x)\right), \tag{3.1}$$

where $\theta(x) = (\sigma(x), \xi(x))$, and the scale and shape are continuous functions $\sigma : \mathcal{X} \to (0, \infty)$ and $\xi : \mathcal{X} \to \mathbb{R}$, respectively. This assumption is a conditional version of (2.1) and means that the GPD approximation holds for the distribution of $Y \mid X = x$ for any $x \in \mathcal{X}$. It is satisfied by most data generating processes as for instance in Example 1.

In order to formulate our estimators of the conditional GPD parameters $\theta(x)$ and the extreme quantile $Q_x(\tau)$, we define the exceedances in the training data as

$$Z_i := (Y_i - \hat{Q}_{X_i}(\tau_0))_+, \quad i = 1, \ldots, n; \tag{3.2}$$

see the triangles in Figure 1. Here, $\tau_0 \in (0, 1)$ is an intermediate probability level that is chosen such that the estimator $\hat{Q}_x(\tau_0)$ of the conditional quantile function can be obtained by classical quantile regression techniques; see Section 2.2. In principle, any quantile regression method can be used to fit $Q_x(\tau_0)$. Here, we choose GRF with quantile loss (Athey et al., 2019) since it is a flexible method well-suited for high-dimensional quantile regression problems and it requires little tuning.

For the estimation of the GPD parameter vector $\theta(x) = (\sigma(x), \xi(x))$ we rely on those exceedances that carry most information on the tail of the distribution of $Y \mid X = x$. To do so, we use the localizing weight functions $w_n(x, X_i)$ estimated from a GRF (Athey et al., 2019) that may be *different* from the one used to estimate the intermediate quantile $\hat{Q}_x(\tau_0)$. We would like to define the estimator of the conditional GPD parameter $\hat{\theta}(x)$ as the minimizer of the weighted (negative) log-likelihood

$$L_n(\theta; x) = \sum_{i=1}^{n} w_n(x, X_i) \ell_\theta(Z_i) 1\{Z_i > 0\}, \quad x \in \mathcal{X}, \tag{3.3}$$

where ℓ_θ is defined in (2.4). In practice, the parameter space $\theta(\mathcal{X}) = \{\vartheta \in (0, \infty) \times \mathbb{R} : \vartheta = \theta(x) \text{ for some } x \in \mathcal{X}\}$ is unknown. As explained by Dombry (2015), it is not guaranteed that the log-likelihood of the generalized extreme value distribution has a global optimum over the parameter space $(0, \infty) \times \mathbb{R}$. In fact, Smith (1985) shows that there exists no maximum likelihood estimator when $\xi \leq -1$. Analogous results apply to the GPD log-likelihood $L_n(\theta; x)$ (Drees et al., 2004). We therefore follow Bücher and Segers (2017) and define $\hat{\theta}(x)$ as the optimizer of $L_n(\theta; x)$ over an arbitrarily large compact set $\Theta \subset (0, \infty) \times (-1, \infty)$ such that $\theta(\mathcal{X}) \subset \operatorname{Int} \Theta$, that is,

$$\hat{\theta}(x) = \operatorname*{arg\,min}_{\theta \in \Theta} L_n(\theta; x). \tag{3.4}$$

If there is more than one minimizer, let $\hat{\theta}(x)$ be the smallest with respect to lexicographic order (see Dombry, 2015). The estimated pair $(\hat{Q}_x(\tau_0), \hat{\theta}(x))$ of intermediate quantile and conditional GPD parameters can be plugged into the extrapolation formula (1.1) to obtain an estimate $\hat{Q}_x(\tau)$ of the extreme conditional quantile at level $\tau > \tau_0$.

In Algorithm 1, we describe our prediction method, which we call the extremal random forest (ERF). The algorithm consists of two subroutines, namely ERF-Fit and ERF-Predict. The first one estimates a similarity weight function $(x, y) \mapsto w_n(x, y)$ and an intermediate quantile function $x \mapsto \hat{Q}_x(\tau_0)$ from the training data, for $x, y \in \mathcal{X}$. The second procedure predicts the extreme τ-quantile $\hat{Q}_x(\tau_0)$ at point $x \in \mathcal{X}$ by estimating the GPD parameter vector $\theta(x)$ as in (3.4). Appendix C shows the estimated GRF weights $w_n(x, X_i)$ used in the likelihood in (3.3) for Example 1 and specific values of x. It can be seen that the weights are large for training observations X_i where the distribution of $Y \mid X = X_i$ is equal to the one of $Y \mid X = x$.

Algorithm 1 Extremal random forest (ERF)

Denote by $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^n$ the training data. Let $x \in \mathbb{R}^p$ be a test predictor value. Specify intermediate and extreme quantile levels τ_0 and τ, respectively, with $\tau_0 < \tau$. Let α be a vector of hyperparameters supplied to GRF.

procedure ERF-Fit($\mathcal{D}, \tau_0, \alpha$)
  $w_n(\cdot,\cdot) \leftarrow \text{GRF}(\mathcal{D}, \alpha)$ ▷ Fit similarity weight function.
  $\hat{Q}_\cdot(\tau_0) \leftarrow$ QuantileRegression($\mathcal{D}$) ▷ Fit intermediate quantile function.
  output `erf` $\leftarrow [\mathcal{D}, w_n(\cdot,\cdot), \hat{Q}_\cdot(\tau_0)]$ ▷ Return an `erf` object.

procedure ERF-Predict(`erf`, x, τ)
  $Z_i \leftarrow (Y_i - \hat{Q}_{X_i}(\tau_0))_+$, with $i = 1, \dots, n$ ▷ Compute exceedances.
  $\hat{\theta}(x) \leftarrow \arg\min_\theta L_n(\theta; x)$ as in (3.3) ▷ Estimate GPD parameters.
  output $\hat{Q}_x(\tau)$ as in (1.1) ▷ Return extreme quantile.

The subroutine GRF estimates the similarity weight function $w_n(\cdot,\cdot)$ using the generalized random forest of Athey et al. (2019). The subroutine QuantileRegression fits the intermediate conditional quantile function $\hat{Q}_\cdot(\tau_0)$ using a classical quantile regression technique. The object `erf` is a list containing the training data $\mathcal{D}$, the fitted intermediate quantile $\hat{Q}_\cdot(\tau_0)$, and the estimated similarity weight function $w_n(\cdot,\cdot)$.

3.2 Consistency

Our ERF provides an estimate $\hat{\theta}(x) = (\hat{\sigma}(x), \hat{\xi}(x))$ of the conditional GPD parameter $\theta(x)$ that describes the tail of the distribution of $Y \mid X = x$. The method is at the interface of random forests and extreme value theory, and both fields have their challenges related to the analysis of asymptotic properties.

Consistency and asymptotic normality of classical (Meinshausen, 2006; Biau, 2012; Scornet et al., 2015; Wager and Athey, 2018) and generalized random forests (Athey et al., 2019) have only recently been established. The results by Athey et al. (2019) require regularity conditions (see Assumptions 1–6 of their paper) that are not satisfied in our setting. In particular, the negative GPD log-likelihood $\theta \mapsto \ell_\theta(z)$ that we consider is not a convex function and, therefore, it does not satisfy Assumption 6 in Athey et al. (2019). On the other hand, the asymptotic analysis of extreme value estimators is notoriously difficult due to the pre-limit approximation in (3.1) and changing distributional support (Smith, 1985; Drees et al., 2004). Recent papers have worked out the asymptotics for the unconditional i.i.d. case (Dombry, 2015; Bücher and Segers, 2017; Dombry and Ferreira, 2019).

We will not show consistency of the ERF under the most general conditions on the distributional tail of $Y \mid X = x$ since the required technicalities would be beyond the scope of this paper. We list all assumptions needed for our theorem and discuss possible relaxations after the statement. The first assumption deals with the data generating process.

Assumption 1. Let $X \in \mathcal{X}$ have a density that is bounded away from 0 and ∞ and support $\mathcal{X} := [0,1]^p$. For large enough τ_0, suppose the conditional intermediate quantile function $Q_X(\tau_0)$ is known. Furthermore, assume that the distribution function of $Y - Q_X(\tau_0)$, conditional on

$Y > Q_X(\tau_0)$, is *exactly* generalized Pareto with parameter vector $\theta(X)$.

The next assumption addresses how the parameter vector $\theta(x)$ depends on the predictor $X = x$. We consider only the most relevant case of positive shape parameter $\xi(x) > 0$, that is, where $Y \mid X = x$ is heavy-tailed.

Assumption 2. Let $\theta(x) = (\sigma(x), \xi(x))$ denote the bivariate regression function for the GPD parameters, for $x \in \mathcal{X}$. Assume $\sigma : \mathcal{X} \to (0, \infty)$ and $\xi : \mathcal{X} \to (0, \infty)$ are continuous functions on $\mathcal{X}$. Furthermore, assume their first order partial derivatives are continuous in the interior and exist on the boundary of $\mathcal{X}$; we refer to Appendix B for a definition of partial derivative on the boundary. Notice that the parameter space $\theta(\mathcal{X}) \subset (0, \infty) \times (0, \infty)$ is compact and bounded away from the origin.

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be independent copies of (X, Y). The first step of our algorithm consists of fitting a generalized random forest on the training data to obtain similarity weights. To show consistency, we make the following standard assumptions on how this forest is built.

Assumption 3. Let $w_n(x, y)$ denote the similarity weights for $x, y \in \mathcal{X}$ estimated by a GRF. We assume the forest satisfies Specification 1 of Athey et al. (2019). In particular, we assume that each tree in the forest is symmetric, places balanced splits, and is randomized (see Athey et al., 2019). We require that each tree is fitted on a subsample of the training data with size $s < n$, such that $s \to \infty$ and $s/n \to 0$ as $n \to \infty$, and that the forest consists of $\binom{n}{s}$ trees fitted on all possible subsamples of size s.

In practice, one builds a forest by estimating B trees. Our theoretical results hold for forests made of $\binom{n}{s}$ trees fitted on all possible subsamples of size s. For this reason, similarly to Wager and Athey (2018), we assume that B is large enough so that the Monte Carlo effect is negligible. Furthermore, Assumption 3 does not require that the trees in the forest are honest in the sense of Athey et al. (2019). The reason is that, as opposed to Athey et al. (2019), our conditional response distribution belongs to the parametric GPD family. In practice, we find that honesty helps our algorithm perform better, and the result below remains true under this additional, stronger assumption.

Theorem 1. *Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be independent copies of (X, Y) as specified in Assumptions 1 and 2. Let $x \in \operatorname{Int} \mathcal{X}$ be a fixed test predictor value, and denote by $w_n(x, X_i)$ the similarity weights estimated with a forest satisfying Assumption 3. Let $\Theta \subset (0, \infty) \times (0, \infty)$ be an arbitrary compact set such that $\theta(\mathcal{X}) \subset \operatorname{Int}(\Theta)$, and let $\hat{\theta}(x)$ denote a sequence of estimators minimizing (3.3). Then, $\hat{\theta}(x) \to \theta(x)$ in probability as $n \to \infty$.*

The proof relies on the theory of Newey (1991) and is in Appendix A. To the best of our knowledge, this is the first consistency proof for a tree-based extreme quantile regression method that works for high-dimensional predictor spaces and complex parameter response surfaces. Wang and Tsai (2009) show asymptotic normality for the model parameters for the heavy-tailed case, but only in the situation where the covariate dependence is linear. There are no asymptotic results for models for generalized Pareto distributions with parameters depending in a more complex way on

the covariates such as through generalized additive models (Chavez-Demoulin and Davison, 2005; Youngman, 2019), trees (Farkas et al., 2020) or gradient boosting (Velthoen et al., 2021).

Similarly to Wang and Tsai (2009), we focus on the heavy-tailed case where $\xi(x) > 0$ for all $x \in \mathcal{X}$. Relaxing this assumption to $\xi(x) \in \mathbb{R}$ would make the support of the generalized Pareto distribution depend on the model parameters. This would require a different proof strategy and additional care in terms of Lipschitz conditions, but some ideas from the i.i.d. case in Bücher and Segers (2017, Lemma E.2) might be helpful.

A further simplification in our setup is that we assume that the approximation in (3.1) is an equality. Dropping this assumption would require additional conditions to control the approximation error and would add a further level of technicality to the proofs. Similar assumptions are often made in the literature as for instance in Bücher and Segers (2017) for the i.i.d. case for generalized extreme value distributions.

3.3 Hyperparameter Tuning

Generalized random forests have several tuning parameters, such as the number of predictors selected at each split and the minimum node size. This section presents a cross-validation scheme to tune such hyperparameters within our algorithm. For large values of $\tau \approx 1$, the quantile loss is not a reliable scoring function since there might be few or no test observations above this level. In our case, we can instead rely on the tail approximation in (3.1) and use the deviance of the GPD as a reasonable metric for cross-validation. Let $\mathcal{N}_1, \dots, \mathcal{N}_M$ be a random partitioning of $\{1, \dots, n\}$ into M equally sized folds of the training data. For a sequence $\alpha_1, \dots, \alpha_J$ of tuning parameters, we fit an `erf` object on the training set (X_i, Y_i), $i \notin \mathcal{N}_m$, for each α_j and each fold m as described in the ERF-Fit function in Algorithm 1. Given the fitted `erf` object, we estimate the GPD parameter vector $\hat{\theta}(X_i; \alpha_j)$ on the validation set (X_i, Y_i), $i \in \mathcal{N}_m$ as in the ERF-Predict function in Algorithm 1, and evaluate the cross-validation error by

$$CV(\alpha_j) = \sum_{m=1}^{M} \sum_{i \in \mathcal{N}_m} \ell_{\hat{\theta}(X_i;\alpha_j)}(Z_i) 1\{Z_i > 0\},$$

where $\theta \mapsto \ell_\theta(z)$ is the deviance of the GPD and $Z_i := (Y_i - \hat{Q}_{X_i}(\tau_0))_+$ are the exceedances. Finally, we select the optimal tuning parameter α^* as the minimizer of $CV(\alpha_j)$, $j = 1, \dots, J$. To make the problem computationally tractable, we first fit the intermediate quantile function $x \mapsto \hat{Q}_x(\tau_0)$ on the entire data set. Then, on each fold, we estimate the similarity weight function $(x, y) \mapsto w_n(x, y)$ with "small" forests made of 50 trees. We repeat the cross-validation scheme several times to reduce the variability of the results.

Even though, in principle, one could perform cross-validation on several tuning parameters, we find that the minimum node size $\kappa \in \mathbb{N}$ plays the most critical role for ERF. The reason is that κ controls the model complexity of the individual trees in the forest and consequently of the similarity weights $w_n(\cdot, \cdot)$. Small (large) values of κ correspond to trees with few (many) observations in each leaf and produce strongly (weakly) localized weight functions $w_n(\cdot, \cdot)$. The estimates of the shape parameter $\hat{\xi}(x)$ in (3.4) may be sensitive to small changes of the localizing weights in the covariate space, leading to unstable quantile predictions through (1.1). To reduce the variance of

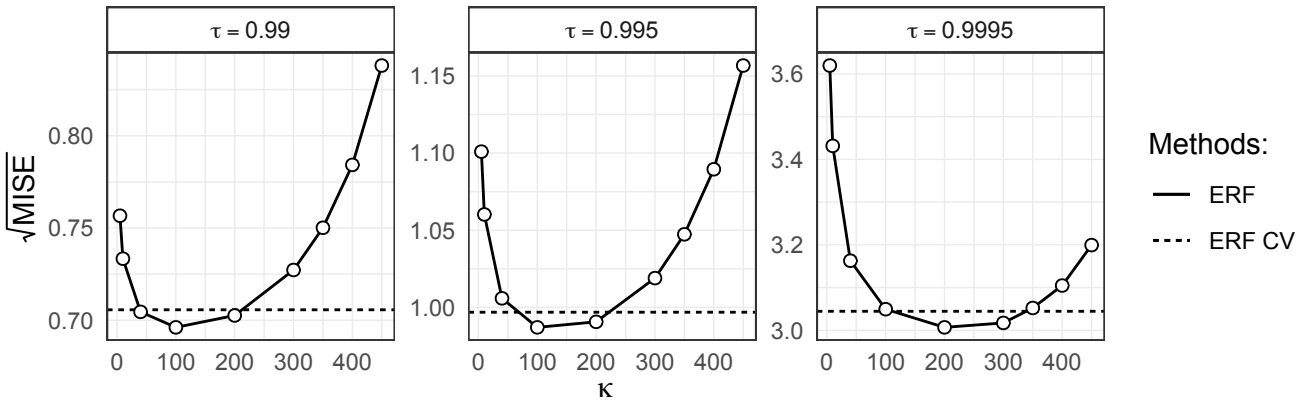

Figure 2: Solid line shows the square root of the MISE of ERF for different minimum node sizes κ over 50 simulations. The dashed line shows the square root MISE of the cross-validated ERF. The data is generated according to Example 1.

$\hat{\xi}(x)$, it is useful to stabilize the log-likelihood $x \mapsto L_n(\theta; x)$ by estimating the similarity weights $w_n(\cdot, \cdot)$ with a forest made of trees with relatively large leaves. Notice that $w_n(x, y)$ influences the effective number of observations used in the weighted (negative) log-likelihood $L_n(\theta; x)$ (3.3).

Figure 2 shows numerical results of cross-validating the minimum node size κ for the model described in Example 1. Here, we perform 5-fold cross-validation repeated three times by growing forests of 50 trees on each fold. We measure the performance as the square root of the mean integrated squared error (MISE) between the estimated and the true quantile function over 50 simulations; see Section 4 for the definition of the MISE. We observe that the cross-validated performance of ERF (dashed line) is close to the minimum square root MISE, suggesting that the proposed cross-validation scheme works well.

3.4 Penalized Log-Likelihood

The shape ξ of the GPD is the most crucial parameter since it determines the tail behavior of Y at extreme quantile levels; the extrapolation formula (2.3) shows the highly non-linear influence of the shape parameter on large quantiles.

Estimation of the shape parameter is notoriously challenging, and the maximization of the GPD likelihood may exhibit convergence problems for small sample sizes (Coles and Dixon, 1999). In general, penalization can help to reduce the variance of an estimator at the cost of higher bias (Hastie et al., 2009). Coles and Dixon (1999) propose a penalty function that restricts the shape parameter values to $\xi < 1$ and favors smaller values of ξ. Several penalization schemes can be interpreted in a Bayesian sense by considering a prior distribution on the regularized parameter. For example, de Zea Bermudez and Turkman (2003) introduce a Bayesian approach to estimate the ξ by using different priors for the cases $\xi > 0$ and $\xi < 0$, respectively. In the context of the generalized extreme value distribution, other penalization methods have been proposed by Smith and Naylor (1987).

While the above regularization methods are tailored to i.i.d. data, in our setting we want to

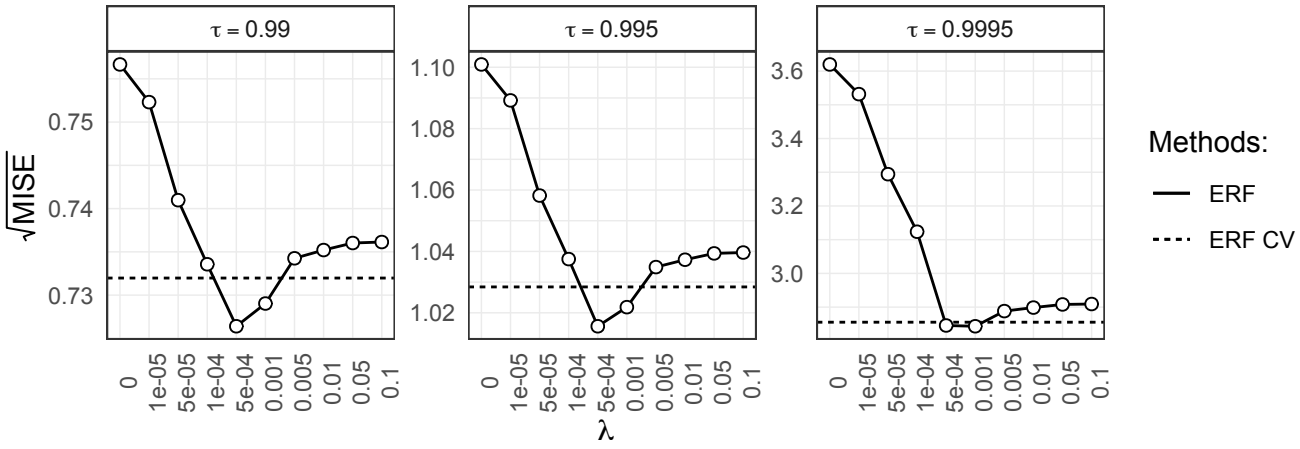

Figure 3: Square root MISE of ERF for different penalty values λ and quantile levels τ over 50 simulations. The data is generated according to Example 1.

penalize the variation of the shape function $x \mapsto \xi(x)$ across the predictor space $\mathcal{X}$. In spatial applications, for instance, it is common to assume a constant shape parameter at different locations (e.g., Ferreira et al., 2012; Engelke et al., 2019). Similarly, in ERF we shrink the estimates $\hat{\xi}(x)$ to a constant shape parameter ξ_0. In general, ξ_0 can be given by expert knowledge, but often a good choice is the unconditional fit $\xi_0 = \hat{\xi}$ obtained by minimizing the GPD deviance in (3.3) with constant weights $w_n(x, y) = 1$ for all $x, y \in \mathcal{X}$.

We propose to penalize the weighted GPD deviance (3.3) with the squared distance between the estimates of $\xi(x)$ and the constant shape parameter ξ_0, that is,

$$\hat{\theta}(x) = \underset{(\sigma,\xi)=\theta\in\Theta}{\arg\min} \frac{1}{(1-\tau_0)} L_n(\theta; x) + \lambda(\xi - \xi_0)^2, \tag{3.5}$$

where $\lambda \geq 0$ is a tuning parameter, and τ_0 is the intermediate quantile level. The parameter λ allows interpolating between a simpler model with a constant shape when $\lambda \to \infty$, and a more complex model with a varying shape over the predictor space when λ is small. This penalized negative log-likelihood can be interpreted in a Bayesian sense: it is equivalent to the maximum *a posteriori* GPD estimator when putting Gaussian prior $N(\xi_0, 1/(2\lambda))$ on the shape parameter ξ. Bücher et al. (2020) propose the same penalization as in (3.5) to estimate the generalized extreme value distribution parameters, where the prior distribution is centered around an expert belief ξ_0 and $\lambda \geq 0$ reflects the confidence in such belief.

In practice, when we penalize the shape parameter we modify Algorithm 1 by replacing Line 3 of the ERF-Predict subroutine with (3.5). Similarly, we cross-validate λ using the scheme presented in Section 3.3 on the modified Algorithm 1. Figure 3 shows the square root MISE over 50 simulations for different values of λ and different quantile levels. Here, we set ξ_0 as the estimated unconditional shape parameter.

4 Simulation Study

4.1 Setup

We compare ERF to other quantile regression methods on simulated data sets, assessing the properties of the different approaches. In the three experiments, we simulate n training observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ as independent copies of a random vector (X, Y). We always generate the predictor $X \in \mathbb{R}^p$ from a uniform distribution on the cube $[-1, 1]^p$ for different dimensions p. We let the conditional response variable $Y \mid X = x$ follow distributions such as Gaussian or Student's t, with tail heaviness depending on the simulation study. The parameters of these distributions, and therefore also the parameters of the GPD corresponding to their tails, vary as functions of the predictor value x. Different response surfaces are considered. The goal is to predict the quantiles $Q_x(\tau)$ of the conditional response $Y \mid X = x$ for moderately to very extreme quantile levels $\tau > 0$.

We evaluate the performance of the method on a test data set $\{x_i\}_{i=1}^{n'}$ of $n' = 1000$ observations generated with a Halton sequence (Halton, 1964) on the cube $[-1, 1]^p$. For a fitted quantile regression function $x \mapsto \hat{Q}_x(\tau)$, $\tau \in (0, 1)$, we then compute the integrated squared error (ISE) on the test data set as

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

where $x \mapsto Q_x(\tau)$ is the true quantile function of the model. Repeating the simulation, fitting and evaluation $m = 50$ times, we obtain the mean integrated squared error (MISE) as the average of the different ISEs.

In the first experiment, we study how ERF performs on the two challenges of high quantile levels and high-dimensional predictor spaces illustrated in Figure 1. The data sets follow the model of Example 1 where the response has a Student's t-distribution with scale shift according to a step function. We consider the methods' performances for different dimensions p of the predictor space and different quantile levels τ.

The second experiment studies the robustness of ERF and other methods to the tail heaviness of the noise distribution. The data generating function is the same as in the first experiment, except that the tail of the noise ranges from the light-tailed Gaussian case with $\xi = 0$ to the relatively heavy tails of Student's t distributions with large $\xi > 0$.

In the last experiment (see Appendix D.2), we consider more complex regression functions for the conditional response variables to assess the performance of the quantile regression methods on complex data. The underlying models depend on more than one predictor value, and both the scale and the shape parameters vary simultaneously.

4.2 Competing Methods

Among the forest-based algorithms, we consider the quantile regression forest by Meinshausen (2006), denoted by QRF, and the generalized random forest by Athey et al. (2019), denoted by GRF. Since these methods do not rely on the GPD likelihood, it is not possible to cross-validate their tuning parameters as in Section 3.3 for prediction error of extreme quantiles. However, in

independent simulations, we notice that their tuning parameters do not have big influence on the results. We set their tuning parameters to the default values and and fit the quantile functions $\hat{Q}_x^{QRF}(\tau)$, $\hat{Q}_x^{GRF}(\tau)$ on the training data for some $\tau \in (0,1)$. More details on forest-based approaches can be found in Section 2.2.

As a hybrid method that uses forest-based weights, we consider the method EGP Tail proposed by Taillardat et al. (2019) who assume that the entire conditional distribution $Y \mid X = x$ follows a parametric family called extended generalized Pareto (EGP) distribution. They estimate the covariate dependent parameters of the EGP through a probability-weighted method of moments using the estimated quantiles $\hat{Q}_x^{GRF}(\tau)$ of the GRF. We follow the authors' implementation and use the default parameter values.

Our ERF method is part of the class of extrapolation approaches that model the exceedances Z_i in (3.2) by conditional GPD distributions. Among the numerous methods that follow this strategy we present only those from Youngman (2019) and Velthoen et al. (2021) as they turn out to be most competitive. Other existing extrapolation based methods are not flexible enough in our setting (Wang and Tsai, 2009; Wang et al., 2012) or do not perform well with larger noise dimensions (Daouia et al., 2011; Gardes and Stupfler, 2019). For the sake of comparability, for all extrapolation methods we use the same exceedances $Z_i = (Y_i - \hat{Q}_x^{GRF}(\tau_0))_+$, which are computed from a GRF with intermediate quantile level $\tau_0 = 0.8 \leq \tau$. To assess the sensitivity of our method to the intermediate threshold τ_0, we perform a simulation study for a data set generated according to Example 1; see Figure 10 in Appendix D.1. In this setup, the intermediate threshold does not strongly influence the results. In general, the optimal choice will depend on the properties of the data (see de Haan and Ferreira, 2006, Section 3.2) and numerous data-driven methods for choosing the threshold exist (e.g., Embrechts et al., 2012, Section 6.2.2).

The method from Youngman (2019) uses generalized additive models to estimate the parameters of a GPD distribution. Here, we model the scale and shape parameters as smooth additive functions of the covariates without interaction effects. In the sequel, we abbreviate this method by EGAM. Velthoen et al. (2021) propose the GBEX method to estimate the GPD parameters using gradient boosting (Friedman, 2001, 2002). In particular, they grow two sequences of gradient trees to model the conditional scale and shape parameter, respectively. To fit GBEX, we use 5-fold cross validation with a maximum number of trees per fold set to $B_{\max} = 500$. We set the depth of each gradient tree $D = 2$, and we set the learning rate for the scale parameter to $\lambda^\sigma = 0.1$. We set the other tuning parameters to their default values. We also consider the unconditional model as a baseline, where we fit constant GPD parameters (σ, ξ) to the conditional exceedances Z_i.

Concerning our ERF method, we fit the parameters as described in Algorithm 1 using the repeated cross-validation scheme described in Section 3.3. In particular we repeat three times 5-fold cross-validation to tune the minimum node size $\kappa \in \{10, 40, 100\}$ and the penalty $\lambda \in \{0, 0.01, 0.001\}$ for the shape parameter. We leave the other tuning parameters of the random forests at their default values; see the documentation for `quantile_forest` in Tibshirani et al. (2021). All simulation results can be reproduced following the description and code on https://github.com/nicolagnecco/erf-numerical-results.

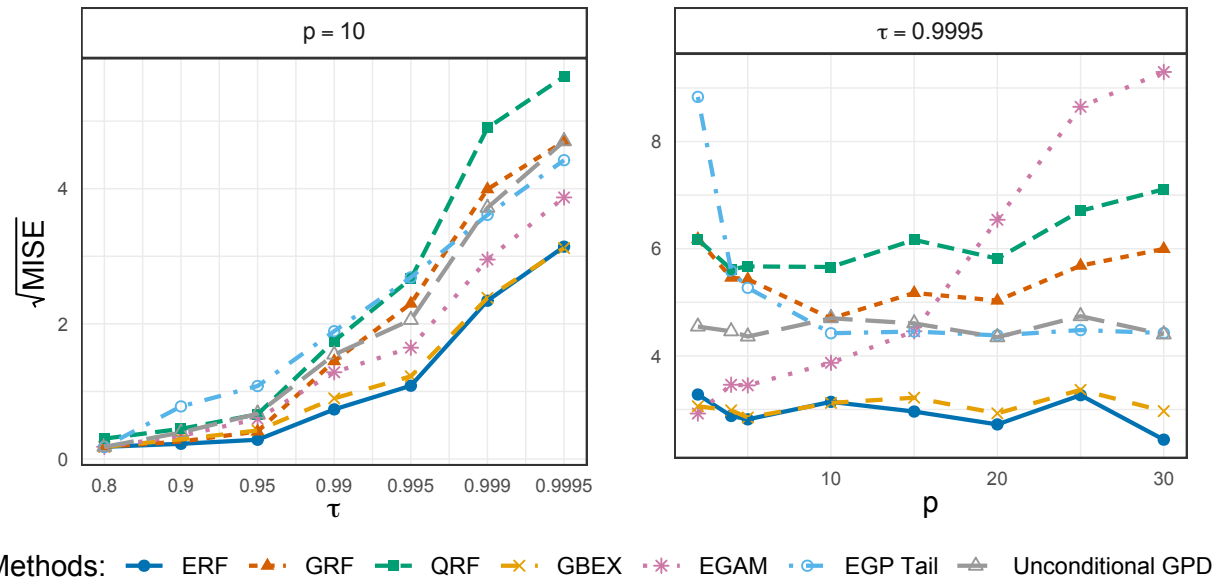

Figure 4: Square root MISE for different methods against the quantile level τ in dimension $p = 10$ (left), and against the model dimension p for quantile levels $\tau = 0.9995$ (right).

4.3 Experiment 1

In this simulation study, the data follows the model of Example 1 where the response variable $Y \mid X = x$ follows a Student's t-distribution with $\nu(x) \equiv 1/\xi(x) = 4$ degrees of freedom and scale $s(x) = 1 + \mathbb{1}\{x_1 > 0\}$. This is the same setup as in the simulation in Athey et al. (2019, Section 5), except that here we use Student's t-distribution instead of Gaussian for the noise. There is only one signal variable X_1 and $p - 1$ noise variables. We generate $n = 2000$ training data and consider different dimensions p and quantile levels τ.

We first fix the dimension $p = 10$ and investigate the effect of different target quantile levels τ on the prediction performances of the competing methods. The left panel of Figure 4 shows the $\sqrt{\text{MISE}}$, the square root of the MISE defined in Section 4.1, for varying values of τ close to 1. At the intermediate quantile level $\tau_0 = 0.8$ all methods show a similar performance; in fact, the extrapolation methods coincide at this level since they use the same GRF based estimator for the intermediate quantile. When the quantile level τ increases, or equivalently, the expected number of exceedances $n(1 - \tau)$ in the training sample decreases, we observe that the performance curves diverge. The forest-based quantile regression methods that do not explicitly use extreme value theory for tail approximations cannot extrapolate well to extreme quantile levels. This includes the EGP Tail method that does not focus on modeling the tail. Among the extrapolation methods, the unconditional baseline does not perform well since it cannot capture the shift in the scale function. While the EGAM does better, it already suffers from the relatively high dimension of the noise variables, a fact that we discuss in detail below. By far, the best methods are our ERF and the GBEX. Both combine the flexibility in the predictor space with correct extrapolation originating from the GPD approximation.

We next compare the performances for varying dimensions p of the predictor space. The right

panel of Figure 4 shows the $\sqrt{\text{MISE}}$ as a function of p for fixed quantile levels $\tau = 0.9995$. QRF and GRF look relatively robust against growing dimensions and additional noise variables, but the performance is not competitive for higher quantiles levels. For smaller dimensions, the methods deteriorate because of the overfitting; the trees can only place split on the signal variable X_1, increasing the variance. The performance of EGAM clearly illustrates the problem of this method with higher dimensions. The method cannot filter the signal from the many noise variables even though, in principle, it is flexible enough to model the response function; the latter is indicated by the good performance for very small noise dimension. Moreover, as mentioned by Youngman (2019), the method becomes computationally demanding as p grows. The unconditional model is unaffected by the noise dimension since it does not use the predictor values. Both ERF and GBEX combine the advantages of the two types of approaches. They are both robust against additional noise variables and perform well even for large dimensional predictor spaces.

4.4 Experiment 2

In the second experiment, we investigate the robustness of the quantile regression methods against noise distributions with different tail heaviness in a large dimension. The simulation setup is similar to the previous section and data follows the model of Example 1, where we set $p = 40$. We simulate data for noise distributions with shape parameters $\xi = 0, 1/4, 1/3$, where for the light-tailed case $\xi = 0$ we choose a Gaussian distribution and otherwise a Student's t distribution with $\xi = 1/4, 1/3$ corresponding $\nu = 4, 3$ degrees of freedom, respectively. We exclude EGAM in this experiment since its performance decreases for large p and it becomes computationally prohibitive (see Figure 4).

Figure 5 shows boxplots of the $\sqrt{\text{ISE}}$ for the extreme quantile level $\tau = 0.9995$ for the different methods and different shape parameters. The triangles correspond to the average values. To make the plot easier to visualize, we remove large outliers of GRF and QRF. The picture is similar for the three noise distributions. We observe that ERF performs very well also in the Gaussian case. Since our method relies on the GPD, estimation is not restricted to positive shape parameters, as opposed to approaches based on the Hill estimator (e.g., Wang et al., 2012; Wang and Li, 2013). Unsurprisingly, as the noise becomes very heavy-tailed (right-hand side of Figure 5) the performances of all methods become closer since the problem becomes increasingly difficult. We further note that the performance of both QRF and GRF degrades for large values of ξ. They exhibit increasingly large outliers that result in an average exceeding the upper quartile. This underlines that classical methods without proper extrapolation are insufficient for extreme quantile regression.

5 Analysis of the U.S. Wage Structure

We compare the performance of ERF, GBEX, GRF, and the unconditional GPD on the U.S. census microdata for the year 1980 (Angrist et al., 2009). As described therein, the data set consists of 65,023 U.S.-born black and white men of age between 40–49, with five to twenty years of education, and with positive annual earnings and hours worked in the year before the census. The large number of observations makes this dataset suitable to assess the performance of the different

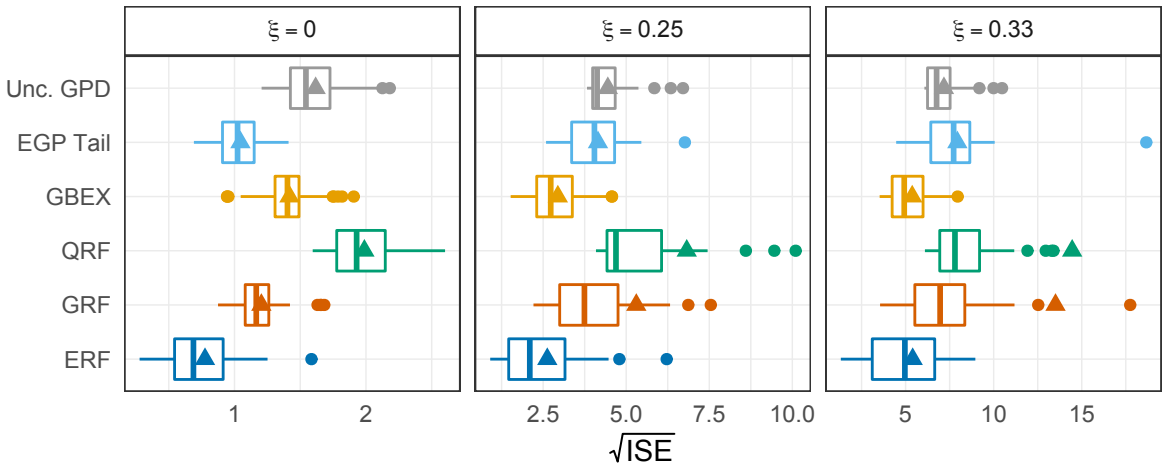

Figure 5: Boxplots of $\sqrt{\text{ISE}}$ over $m = 50$ simulations, for different tail indices in the noise distribution at the quantile level $\tau = 0.9995$. The predictor space dimension is $p = 40$. Triangles represent the average values.

methods at very high quantile levels. The response Y describes the weekly wage, expressed in 1989 U.S. dollars computed as the annual income divided by the number of weeks worked. The predictor vector consists of the numerical variables age and years of education and the categorical predictor whether the person is black or white. To make the data set higher dimensional, we add ten random predictors sampled independently from uniform distributions on the interval $[-1, 1]$, resulting in a predictor space's dimension $p = 13$.

Throughout this analysis, we fit ERF repeating three times 5-fold cross-validation to tune the minimum node size $\kappa \in \{5, 40, 100\}$. To stabilize the variance of the shape parameter, we set the penalty $\lambda = 0.01$. Regarding the other methods, we use the same tuning parameter setup as in 4.2. In particular, we use GRF to predict the intermediate conditional quantiles at level $\tau_0 = 0.8$ for all extrapolation-based methods. We split the original data into two halves, i.e., 32,511 and 32,512 samples, respectively. We use the first portion to perform an exploratory data analysis and the second one to fit and evaluate the different methods.

For the exploratory data analysis, we fit ERF on a random subset made of 10% of the data (i.e., 3,251 observations), and predict the GPD parameters $\hat{\theta}(x) = (\hat{\sigma}(x), \hat{\xi}(x))$ on the left-out observations (i.e., 29,260 observations). Figure 6 shows the estimated GPD parameters $\hat{\theta}(x)$ as a function of years of education. We observe that the scale parameter depends positively on years of education, whereas it is quite homogeneous between the black and white groups. In particular, it has a clear jump around 15-16 years of education, which corresponds to the end of the undergraduate studies. The shape parameter is relatively homogeneous for the black and white group and looks stable for education. It ranges between 0.22 and 0.24, indicating heavy-tails throughout the predictor space. Moreover, Figure 12 in Appendix E.1 shows that the scale and shape parameters do not seem to depend on the predictor age.

In Figure 7 we compare the ERF quantile predictions to the ones obtained by the other methods at levels $\tau = 0.9, 0.995$. To help with the visualization, we removed all the quantiles above 6,000

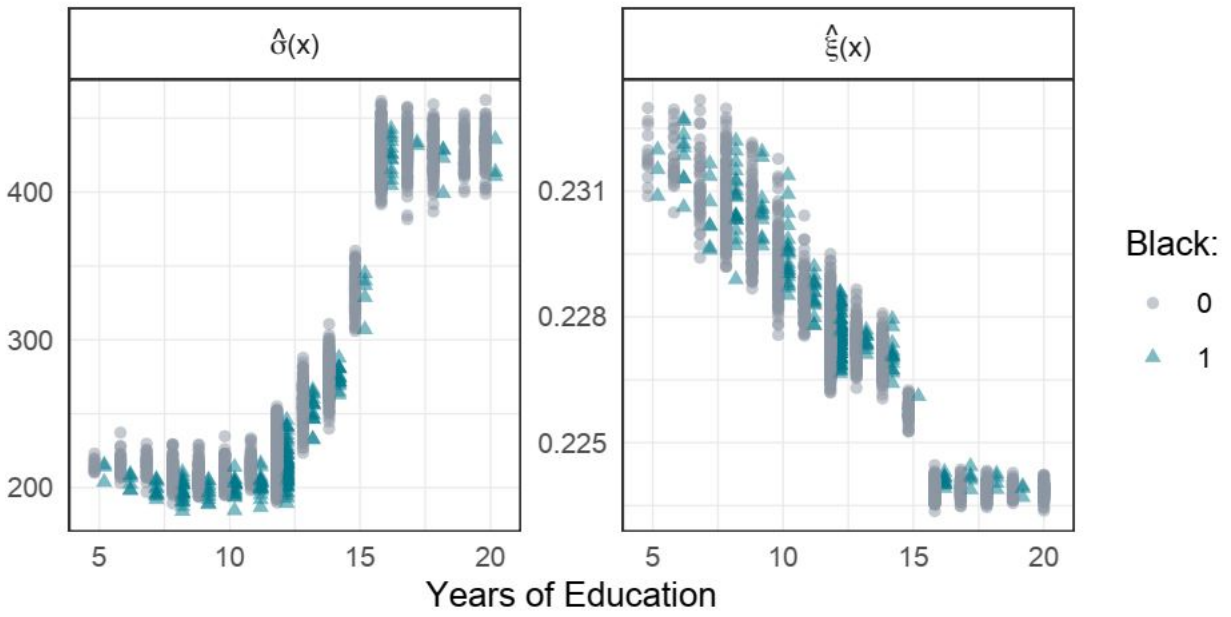

Figure 6: Estimated GPD parameters $\hat{\theta}(x)$ as a function of the years of education for the black (triangles) and white (circles) subgroups.

predicted by GRF. We observe that the extrapolation methods retain a good shape of the quantile function even for high levels. This does not hold for GRF, whose profile worsens as τ increases, and the discrete structure of the largest training observations becomes visible. The unconditional method seems to capture the variability of the conditional quantiles for $\tau = 0.9$, but we observe that it loses flexibility for larger values of τ. The reason for this is that the unconditional method cannot produce different scale parameters of the GPD, while Figure 6 indicates that this is necessary for this data set. ERF and GBEX model well the variability of the conditional quantiles for all values of τ, and they agree on the magnitude of the estimates.

After the exploratory analysis, we assess the quantitative performance of ERF compared to the other methods. We consider the prediction metric proposed by Wang and Li (2013),

$$\mathcal{R}_n\left(\hat{Q}_\cdot(\tau)\right) := \frac{\sum_{i=1}^{n} \mathbb{1}\{Y_i < \hat{Q}_{X_i}(\tau)\} - n\tau}{\sqrt{n\tau(1-\tau)}}, \tag{5.1}$$

where n is the number of test observations, and $\hat{Q}_\cdot(\tau)$ is the τ-th conditional quantile estimated on the training data set. This metric compares the normalized estimated proportion of observations with $Y_i < \hat{Q}_{X_i}(\tau)$ with the theoretical level τ. Using the true quantile function $Q_\cdot(\tau)$, the random variable $\mathbb{1}\{Y_i < Q_{X_i}(\tau)\}$ follows a Bernoulli distribution with expectation τ and variance $\tau(1-\tau)$, and by the central limit theorem the metric with oracle quantile function $\mathcal{R}_n(Q_\cdot(\tau))$ is asymptotically standard normal. We partition the 32,512 observations not used in the exploratory analysis into ten random folds. On each fold, we fit the different methods and evaluate them on the left-out observations, using the absolute value of (5.1). Unlike classical cross-validation, we fit the methods using a single fold and validate them on the remaining ones; this allows us to have enough observations to gauge

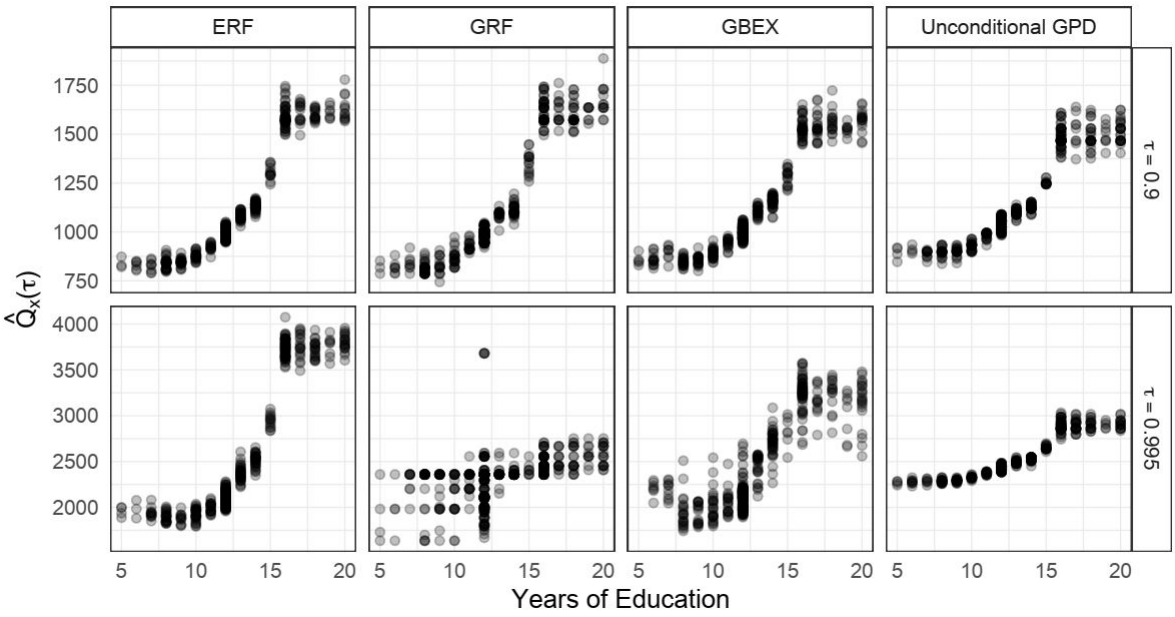

Figure 7: Predicted quantiles at levels $\tau = 0.9, 0.995$ for ERF, GRF, GBEX, and the unconditional method.

their performance for high quantile levels τ. Figure 8 shows the performance of ERF, GRF, GBEX, and the unconditional method over the ten repetitions for different quantile levels. The shaded area represents the 95% interval of the absolute value of a standard normal distribution, corresponding to the 95% confidence level of the oracle method with true quantile function. We observe that both ERF and GBEX have very good performance compared to the oracle for increasing quantile levels, and they outperform the unconditional method for large values of τ. This is because they are flexible to model the scale and shape as a function of the predictors, unlike the unconditional method. While GRF performs well for the quantile level $\tau = 0.9$, it worsens quite quickly for larger values of τ. This is expected since GRF does not rely on extrapolation results from extreme value theory and cannot accurately predict very high quantiles.

For the same data set, Angrist et al. (2006) consider the natural logarithm of the wage as a response variable for quantile regression with fixed, non-extreme quantile levels. In Appendix E.1 we perform our analysis above for extreme quantiles again with this log-transformed response since it highlights several interesting properties of the ERF algorithm. In particular, Figure 14 in Appendix E.2 shows that the flexible methods ERF and GBEX have the desirable property that the predictions do not change much under marginal transformations. The unconditional method, on the other hand, seems to be sensitive to marginal transformations; for an explanation and details, see Appendix E.1. In general, therefore, it is advised to use a flexible extrapolation method, such as ERF or GBEX, that performs well on any marginal distributions.

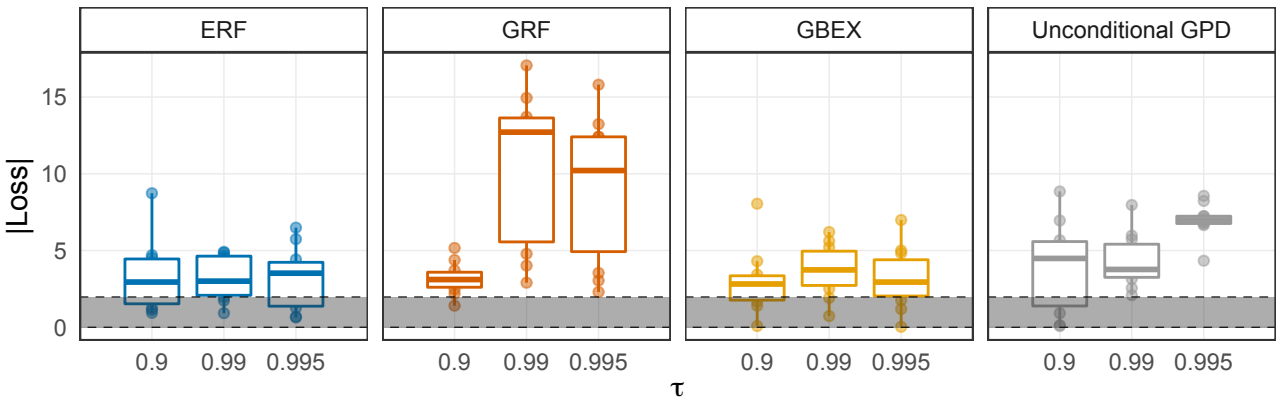

Figure 8: Absolute value of the loss (5.1) for the different methods fitted on the original response of the U.S. wage data. The shaded area represents the 95% interval of the absolute value of a standard normal distribution.

A Proof of Theorem 1

Given the data generating process of Assumption 1 in the main text, define the random variable $Z = (Y - Q_X(\tau_0))_+$. We then have the stochastic representation

$$(X, Z, 1\{Z > 0\}) \overset{d}{=} (X, VP, P), \tag{A.1}$$

where V follows a GPD with parameter vector $\theta(X)$, and $P \sim \text{Bernoulli}(1 - \tau_0)$, independent of X and V. Similarly, for the training data (X_i, Y_i) we may use a an analogous representation with $(X_i, V_i P_i, P_i)$ as in (A.1), $i = 1, \ldots, n$. With this we can rewrite the weighted (negative) log-likelihood function in (3.3) as

$$L_n(\theta; x) = \sum_{i=1}^{n} w_n(x, X_i) \ell_\theta(V_i) P_i,$$

Moreover, for a fixed predictor value $x \in \text{Int}\, \mathcal{X}$ let V^* denote a GPD with parameter vector $\theta(x)$ and define $L(\theta; x) = \mathbb{E}[\ell_\theta(V^*)P]$, where $\theta \in (0, \infty)^2$. To prove our result we rely on Theorem 5.7 of van der Vaart (1998), which we state here adapted to our setting.

Theorem 2. *Let $\theta \mapsto L_n(\theta; x)$ be random functions, and let $\theta \mapsto L(\theta; x)$ be a fixed function such that, for $x \in \text{Int}\, \mathcal{X}$, it holds*

$$\sup_{\theta \in \Theta} |L_n(\theta; x) - L(\theta; x)| \xrightarrow{\mathbb{P}} 0, \tag{A.2}$$

$$L(\theta(x); x) < \inf \left\{ L(\theta; x) : \left\| \theta - \theta(x) \right\|_2 \geq \delta,\ \theta \in \Theta \right\}, \text{ for all } \delta > 0. \tag{A.3}$$

Then any sequence of estimators $\hat{\theta}(x)$ with $L_n(\hat{\theta}(x); x) \leq L_n(\theta(x); x) + o_P(1)$ converges in probability to $\theta(x)$.

We can now prove our Theorem 1.

Proof of Theorem 1. First, notice that $\theta(x) \in \theta(\mathcal{X}) \subset \Theta$, where Θ is compact. Therefore, from (3.4) in the main text, we have that $L_n(\hat{\theta}(x); x) \leq L_n(\theta(x); x)$ for all $n > 0$. Furthermore, a standard argument using Kullback–Leibler divergence implies the true parameter $\theta(x)$ is a minimizer for $\theta \mapsto L(\theta; x)$. Since the GPD is identifiable, the true parameter is a unique minimizer, satisfying condition (A.3). Moreover, from Lemma 1, condition (A.2) is satisfied.

Therefore, from Theorem 2, the estimator $\hat{\theta}(x) \to \theta(x)$ in probability as $n \to \infty$. □

Lemma 1. *Under the assumptions of Theorem 1, it holds that* $\sup_{\theta\in\Theta}|L_n(\theta; x) - L(\theta; x)| \xrightarrow{\mathbb{P}} 0$.

Proof. We have that

$$\begin{aligned} L_n(\theta; x) &= \sum_{i=1}^{n} w_n(x, X_i)\ell_\theta(V_i)P_i \\ &= \sum_{i=1}^{n} w_n(x, X_i)\ell_\theta(V_i^*)P_i + \sum_{i=1}^{n} w_n(x, X_i)\left(\ell_\theta(V_i) - \ell_\theta(V_i^*)\right)P_i \\ &= S_{1,n}(\theta) + S_{2,n}(\theta), \end{aligned}$$

where we couple the random variables V_i and V_i^* through $V_i = F^{-1}_{\theta(X_i)}(U_i)$, $V_i^* = F^{-1}_{\theta(x)}(U_i)$, $U_i \overset{iid}{\sim} \text{Unif}[0, 1]$, and F^{-1}_θ is the inverse of the GPD function with parameter $\theta \in \Theta$. By Lemma 2 and 6, the claim follows. □

Lemma 2. *Under the assumptions of Theorem 1, it holds that* $\sup_{\theta\in\Theta} |S_{1,n}(\theta) - L(\theta; x)| \xrightarrow{\mathbb{P}} 0$.

Proof. Corollary 2.2 of Newey (1991) provides sufficient conditions for uniform convergence.

1. (Compactness): Θ is compact.
2. (Pointwise convergence): For each $\theta \in \Theta$, $S_{1,n}(\theta) - L(\theta; x) = o_P(1)$.
3. (Stochastic Equicontinuity): There exists $C_n = O_P(1)$ such that for all $\theta, \theta' \in \Theta$, $|S_{1,n}(\theta) - S_{1,n}(\theta')| \leq C_n\|\theta - \theta'\|_2$.
4. (Continuity): The map $\theta \mapsto L(\theta; x)$ is continuous.

Condition 1 holds by assumption. The remaining conditions are shown in Lemmas 3, 4, and 5, respectively. □

Lemma 3. *For each $\theta \in \Theta$, it holds that $S_{1,n}(\theta) - L(\theta; x) = o_P(1)$.*

Proof. For each $\theta \in \Theta$, recall that $L(\theta; x) = \mathbb{E}[\ell_\theta(V^*)P]$, where $V^* \sim \text{GPD}(\theta(x))$. Furthermore, we have that

$$S_{1,n}(\theta) = \sum_{i=1}^{n} w_n(x, X_i)\ell_\theta(V_i^*)P_i = \frac{1}{B}\sum_{b=1}^{B}\sum_{i=1}^{n} w_{n,b}(x, X_i)\ell_\theta(V_i^*)P_i = \frac{1}{B}\sum_{b=1}^{B} T_{n,b}(x, \theta),$$

where $T_{n,b}(x,\theta)$ is the output of a regression gradient tree (Athey et al., 2019) with response $\ell_\theta(V_i^*)P_i$ independent of $w_{n,b}(x,X_i)$, $i=1,\dots,n$. Consider a tree $T_{n,b}(x,\theta)$, $b=1,\dots,B$, of the generalized random forest. Its expectation writes

$$\begin{aligned}\mathbb{E}\left(T_{n,b}(x,\theta)\right) &= \sum_{i=1}^{n}\mathbb{E}\left(w_{n,b}(x,X_i)\ell_\theta(V_i^*)P_i\right) = \sum_{i=1}^{n}\mathbb{E}\left(w_{n,b}(x,X_i)\right)\mathbb{E}\left(\ell_\theta(V_i^*)P_i\right)\\ &= \mathbb{E}\left\{\sum_{i=1}^{n} w_{n,b}(x,X_i)\right\}\mathbb{E}\left(\ell_\theta(V^*)P\right) = \mathbb{E}\left(\ell_\theta(V^*)P\right) = L(\theta;x),\end{aligned} \tag{A.4}$$

since the weights sum to one. Therefore, $\mathbb{E}\left(S_{1,n}(\theta)\right) = L(\theta;x)$. Concerning the variance of the thinning $\ell_\theta(V^*)P$ we have that

$$V\left(\ell_\theta(V^*)P\right) = \mathbb{E}\left(\ell_\theta(V^*)^2\right)\mathbb{E}(P^2) - \mathbb{E}\left(\ell_\theta(V^*)\right)^2\mathbb{E}\left(P\right)^2 < +\infty, \tag{A.5}$$

since P is a Bernoulli variable and $\ell_\theta(V^*)$ has exponential tail. Therefore, the variance of $T_{n,b}(x,\theta)$ writes

$$\begin{aligned}\mathbb{V}\left(T_{n,b}(x,\theta)\right) &= \mathbb{E}\left\{\left(T_{n,b}(x,\theta) - L(\theta;x)\right)^2\right\} = \mathbb{E}\left\{\left(\sum_{i=1}^{n} w_{n,b}(x,X_i)\left(\ell_\theta(V_i^*)P_i - L(\theta;x)\right)\right)^2\right\}\\ &= \mathbb{E}\left(\sum_{i=1}^{n} w_{n,b}(x,X_i)^2\left(\ell_\theta(V_i^*)P_i - L(\theta;x)\right)^2\right.\\ &\quad\left. + \sum_{i\neq j} w_{n,b}(x,X_i)w_{n,b}(x,X_j)(\ell_\theta(V_i^*)P_i - L(\theta;x))(\ell_\theta(V_j^*)P_j - L(\theta;x))\right)\\ &= \mathbb{V}\left(\ell_\theta(V^*)P\right)\mathbb{E}\left(\sum_{i=1}^{n} w_{n,b}(x,X_i)^2\right) \leq \mathbb{V}(\ell_\theta(V^*)P) < +\infty,\end{aligned} \tag{A.6}$$

where the fourth equality holds because (V_i^*, P_i) are i.i.d. , the second last inequality holds because $0 \leq w_{n,b}(x,X_i) \leq 1$, and the last inequality follows from (A.5). Using results about U-statistics (Hoeffding, 1948), Wager and Athey (2018) show that the variance of a forest is at most s/n times the variance of a tree, that is

$$\limsup_{n\to\infty} \frac{n}{s}\frac{\mathbb{V}\left(S_{1,n}(\theta)\right)}{\mathbb{V}\left(T_{n,b}(x,\theta)\right)} \leq 1. \tag{A.7}$$

where $s < n$ denotes the subsample size. From Assumption 3, we have that $s/n \to 0$, therefore (A.6) and (A.7) imply that $\mathbb{V}\left(S_{1,n}(\theta)\right) \to 0$ as $n \to 0$. The result follows from Markov's inequality. □

Lemma 4. *There exists $C_n = O_P(1)$ such that for all θ, $\theta' \in \Theta$, $|S_{1,n}(\theta) - S_{1,n}(\theta')| \leq C_n \|\theta - \theta'\|_2$.*

Proof. The negative log-likelihood $\theta \mapsto \ell_\theta(z)$ is defined for each $z \geq 0$ and $\theta \in (0,\infty)^2$ as

$$\ell_\theta(z) = \log\sigma + \left(1 + \frac{1}{\xi}\right)\log\left(1 + \frac{\xi}{\sigma}z\right).$$

Therefore, its partial derivatives can be bounded by

$$\begin{aligned} \left|\partial_\xi \ell_\theta(z)\right| &\leq \frac{1}{\xi^2}\log\left(1 + \frac{\xi}{\sigma}z\right) + \frac{1 + \frac{1}{\xi}}{\xi}, \\ \left|\partial_\sigma \ell_\theta(z)\right| &\leq \frac{1}{\sigma} + \frac{1 + \frac{1}{\xi}}{\sigma}, \end{aligned} \tag{A.8}$$

for any $\theta = (\sigma, \xi) \in (0,\infty)^2$. The bounds from (A.8) are continuous on the compact set $\Theta \subset (0,\infty)^2$, and therefore, from an application of the dominated convergence theorem,

$$g(z) := \sup\left\{\left|\partial_\xi \ell_\theta(z)\right| : \theta \in \Theta\right\} + \sup\left\{\left|\partial_\sigma \ell_\theta(z)\right| : \theta \in \Theta\right\} \tag{A.9}$$

is integrable with respect to a GPD with parameter vector $\theta(x)$. Moreover, for any $\theta, \theta' \in \Theta$, the mean-value theorem and the Cauchy–Schwarz inequality imply

$$\left|\ell_\theta(z) - \ell_{\theta'}(z)\right| = \left|\nabla \ell_{\widetilde{\theta}}(z)(\theta - \theta')\right| \leq \left\|\nabla \ell_{\widetilde{\theta}}(z)\right\|_2 \left\|\theta - \theta'\right\|_2, \tag{A.10}$$

where $\widetilde{\theta} = c\theta + (1-c)\theta'$ for some $0 < c < 1$, and $z \geq 0$. Furthermore, from (A.9), we have that

$$\left\|\nabla \ell_{\widetilde{\theta}}(z)\right\|_2 \leq \left|\partial_\xi \ell_{\widetilde{\theta}}(z)\right| + \left|\partial_\sigma \ell_{\widetilde{\theta}}(z)\right| \leq g(z). \tag{A.11}$$

From equations (A.10) and (A.11) it follows that $\ell_\theta(z)$ is Lipschitz in $\theta \in \Theta$ with constant $g(z)$, $z \geq 0$. Therefore,

$$\begin{aligned} |S_{1,n}(\theta) - S_{1,n}(\theta')| &= \left|\sum_{i=1}^n w_n(x, X_i)\left(\ell_\theta(V_i^*) - \ell_{\theta'}(V_i^*)\right)P_i\right| \leq \sum_{i=1}^n w_n(x, X_i)P_i\left|\ell_\theta(V_i^*) - \ell_{\theta'}(V_i^*)\right| \\ &\leq \left(\sum_{i=1}^n w_n(x, X_i)g(V_i^*)P_i\right)\left\|\theta - \theta'\right\|_2 =: C_n\left\|\theta - \theta'\right\|_2. \end{aligned}$$

For every $n \in \mathbb{N}$ and $i = 1, \ldots, n$, V_i^* is independent of $w_n(x, X_i)$ and P_i. Therefore, since $z \mapsto g(z)$ is integrable with respect to a GPD with parameter vector $\theta(x)$ it follows that $\mathbb{E}[C_n] < +\infty$. Hence, $C_n = O_P(1)$. □

Lemma 5. *The map $\theta \mapsto L(\theta; x)$ is continuous.*

Proof. For any $\theta \in \Theta$, recall that

$$L(\theta; x) = \mathbb{E}[\ell_\theta(V^*)P] = \left\{ \log\sigma + \left(1 + \frac{1}{\xi}\right) \mathbb{E}\left(\log\left[1 + \frac{\xi}{\sigma} V^* \right] \right) \right\} (1 - \tau_0).$$

The maps $\theta \mapsto \log\sigma$ and $\theta \mapsto (1 + 1/\xi)$ are continuous for $\theta \in \Theta$. Also, by an application of the dominated convergence theorem, the map $\theta \mapsto \mathbb{E}\left[\log\left(1 + \frac{\xi}{\sigma}V^*\right)\right]$ is continuous for $\theta \in \Theta$. □

Lemma 6. *Under the assumptions of Theorem 1, it holds that* $\sup_{\theta\in\Theta} |S_{2,n}(\theta)| \xrightarrow{\mathbb{P}} 0$.

Proof. We have that

$$\begin{aligned}
0 \leq \sup_{\theta\in\Theta} |S_{2,n}(\theta)| &= \sup_{\theta\in\Theta} \left| \sum_{i=1}^n w_n(x, X_i) P_i \left(\ell_\theta \circ F^{-1}_{\theta(X_i)}(U_i) - \ell_\theta \circ F^{-1}_{\theta(x)}(U_i) \right) \right| \\
&\leq \sup_{\theta\in\Theta} \sum_{i=1}^n w_n(x, X_i) P_i \left| \ell_\theta \circ F^{-1}_{\theta(X_i)}(U_i) - \ell_\theta \circ F^{-1}_{\theta(x)}(U_i) \right| \\
&\leq \sup_{\theta\in\Theta} \sum_{i=1}^n w_n(x, X_i) P_i K(\theta, U_i) \|X_i - x\|_2 \\
&\leq \sup\left\{ \|X_i - x\|_2 : w_n(x, X_i) > 0, i = 1, \ldots, n \right\} \sum_{i=1}^n w_n(x, X_i) P_i \sup_{\theta\in\Theta} K(\theta, U_i) \\
&= o_P(1),
\end{aligned} \tag{A.12}$$

where the second last inequality follows from Lemma 7.a) and the last equality follows from Lemmas 8 and 7.b). □

Lemma 7. *Let* $x \in \operatorname{Int} \mathcal{X}$, $U \sim \operatorname{Unif}[0, 1]$, *and* $\theta \in \Theta$.

a) *Then, there exists a function* $K(\theta, U) < +\infty$ *such that for any* $y \in \mathcal{X}$,

$$\left| \ell_\theta \circ F^{-1}_{\theta(y)}(U) - \ell_\theta \circ F^{-1}_{\theta(x)}(U) \right| \leq K(\theta, U) \left\| y - x \right\|_2 .$$

b) *Then, under the assumptions of Theorem 1, it holds that*

$$\sum_{i=1}^n w_n(x, X_i) P_i \sup_{\theta\in\Theta} K(\theta, U_i) = O_P(1).$$

Proof.

a) Let $U \sim \operatorname{Unif}[0, 1]$, and $\theta \in \Theta$. For any $y \in \mathcal{X}$ define

$$g(y; \theta, W) := \ell_\theta \circ F^{-1}_{\theta(y)}(1 - 1/W) = \log\sigma + \left(1 + \frac{1}{\xi}\right) \log\left(1 + \frac{\xi}{\sigma} \frac{\sigma(y)}{\xi(y)} \left\{ W^{\xi(y)} - 1 \right\} \right), \tag{A.13}$$

where $W := 1/(1-U) \sim \text{Pareto}(1)$ with support $[1,\infty)$. The map $y \mapsto g(y;\theta,W)$ admits partial derivatives with respect to y_j, $j = 1,\dots,p$, i.e.,

$$\begin{aligned}\partial_{y_j} g(y;\theta,W) = &\left(1+\frac{1}{\xi}\right)\left(1+\frac{\xi}{\sigma}\frac{\sigma(y)}{\xi(y)}\left\{W^{\xi(y)}-1\right\}\right)^{-1}\frac{\xi}{\sigma}\\ &\times\left(\frac{\sigma_j'(y)\xi(y)-\sigma(y)\xi_j'(y)}{\xi(y)^2}\left\{W^{\xi(y)}-1\right\}+\frac{\sigma(y)}{\xi(y)}\left\{W^{\xi(y)}\log W\right\}\xi_j'(y)\right),\end{aligned} \tag{A.14}$$

where σ_j', and ξ_j' are the jth partial derivatives of $y \mapsto \sigma(y)$ and $y \mapsto \xi(y)$, respectively. From Assumption 2 in the main text, we know that $y \mapsto \partial_{y_j} g(y;\theta,W)$ are continuous on the interior of $\mathcal{X}$. Thus, for $x \in \text{Int}\,\mathcal{X}$ and $y \in \mathcal{X}$, the mean-value theorem and the Cauchy–Schwarz inequality imply

$$|g(y;\theta,W) - g(x;\theta,W)| \le \left\|\nabla g(x';\theta,W)\right\|_2 \left\|y-x\right\|_2,$$

where $x' = cy + (1-c)x$ for some $c \in (0,1)$. Moreover, Assumption 2 ensures that the partials derivatives of $y \mapsto g(y;\theta,W)$ exist on the compact set $\mathcal{X}$. Thus, we can define $K(\theta,U) := \sum_{j=1}^p \sup\{|\partial_{y_j} g(y;\theta,W)| : y \in \mathcal{X}\}$ and obtain

$$\left|\ell_\theta \circ F^{-1}_{\theta(y)}(U) - \ell_\theta \circ F^{-1}_{\theta(x)}(U)\right| \le K(\theta,U)\left\|y-x\right\|_2.$$

b) From Part a), we have that $K(\theta,U) = \sum_{j=1}^p \sup\{|\partial_{y_j} g(y;\theta,W)| : y \in \mathcal{X}\}$, where $\theta \in \Theta$, and $W \ge 1$ follows a standard Pareto distribution. For every $j = 1,\dots,p$ it holds that

$$\begin{aligned}\sup_{y\in\mathcal{X}}|\partial_{y_j} g(y;\theta,W)| \le &\sup_{y\in\mathcal{X}}\left(1+\frac{1}{\xi}\right)\left(1+\frac{\xi}{\sigma}\frac{\sigma(y)}{\xi(y)}\left\{W^{\xi(y)}-1\right\}\right)^{-1}\frac{\xi}{\sigma}\\ &\times\left(\frac{|\sigma_j'(y)\xi(y)-\sigma(y)\xi_j'(y)|}{\xi(y)^2}\left\{W^{\xi(y)}-1\right\}+\frac{\sigma(y)}{\xi(y)}|\xi_j'(y)|\left\{W^{\xi(y)}\log W\right\}\right)\\ =: &\sup_{y\in\mathcal{X}}\left(1+\frac{1}{\xi}\right)\frac{\xi}{\sigma}\left(\frac{M_{1j}(y)\left\{W^{\xi(y)}-1\right\}}{1+M(y,\theta)\left\{W^{\xi(y)}-1\right\}}+\frac{M_{2j}(y)\left\{W^{\xi(y)}\log W\right\}}{1+M(y,\theta)\left\{W^{\xi(y)}-1\right\}}\right),\end{aligned}$$

where $M_{1j}(y) = |\sigma_j'(y)\xi(y) - \sigma(y)\xi_j'(y)|/\xi(y)^2 \ge 0$, $M_{2j}(y) = \sigma(y)|\xi_j'(y)|/\xi(y) \ge 0$, and $M(y,\theta) = (\sigma(y)\xi)/(\xi(y)\sigma) > 0$. Notice that almost surely

$$\begin{aligned}0 \le &\frac{M_{1j}(y)\left\{W^{\xi(y)}-1\right\}}{1+M(y,\theta)\left\{W^{\xi(y)}-1\right\}} \le \frac{M_{1j}(y)}{M(y,\theta)},\\ 0 \le &\frac{M_{2j}(y)W^{\xi(y)}}{1+M(y,\theta)\left\{W^{\xi(y)}-1\right\}} \le \max\left\{M_{2j}(y),\frac{M_{2j}(y)}{M(y,\theta)}\right\}.\end{aligned}$$

Therefore, for every $j = 1, \dots, p$, we have

$$\begin{aligned}\sup_{\theta\in\Theta,\ y\in\mathcal{X}} |\partial_{y_j} g(y;\theta,W)| &\le \sup_{\theta\in\Theta,\ y\in\mathcal{X}} \left(1+\frac{1}{\xi}\right)\frac{\xi}{\sigma}\left(\frac{M_{1j}(y)}{M(y,\theta)} + \max\left\{M_{2j}(y), \frac{M_{2j}(y)}{M(y,\theta)}\right\}\log W\right)\\ &\le \left(1+\frac{1}{\xi^-}\right)\frac{\xi^+}{\sigma^-}\left(\frac{M_{1j}}{M} + \max\left\{M_{2j}, \frac{M_{2j}}{M}\right\}\log W\right)\\ &= \left(1+\frac{1}{\xi^-}\right)\frac{\xi^+}{\sigma^-}\left(\frac{M_{1j}}{M} + \frac{M_{2j}}{M}\log W\right),\end{aligned}$$

where $M_{hj} := \sup\{M_{hj}(y) : y \in \mathcal{X}\}$, for $h = 1, 2$, $M := \inf\{M(y,\theta) : \theta \in \Theta,\ y \in \mathcal{X}\} < 1$, and σ^+, ξ^+ (σ^-, ξ^-) are the maxima (minima) of the parameter values over the compact set Θ, respectively. Since $W \sim$ Pareto(1) with support $[1, \infty)$, it follows that $\log W \sim \text{Exp}(1)$. Therefore, by taking expectation we obtain

$$\mathbb{E}\left(\sup_{\theta\in\Theta} K(\theta, U)\right) \le \left(1+\frac{1}{\xi^-}\right)\frac{\xi^+}{\sigma^-}\sum_{j=1}^{p}\left(\frac{M_{1j}+M_{2j}}{M}\right) =: M^* < \infty.$$

Let $\varepsilon > 0$ and consider $M_\varepsilon = (M^* + 1)/\varepsilon > 0$. Then, for any $n \in \mathbb{N}$, it holds that

$$\begin{aligned}&\mathbb{P}\left(\sum_{i=1}^{n} w_n(x, X_i) P_i \sup_{\theta\in\Theta} K(\theta, U_i)) > M_\varepsilon\right) \le \frac{\mathbb{E}\left(\sum_{i=1}^{n} w_n(x, X_i) P_i \sup_{\theta\in\Theta} K(\theta, U_i)\right)}{M_\varepsilon}\\ &= \frac{\sum_{i=1}^{n}\mathbb{E}\left(w_n(x, X_i)\right)\mathbb{E}\left(\sup_{\theta\in\Theta} K(\theta, U_i)\right)\mathbb{E}\left(P_i\right)}{M_\varepsilon} = \frac{\mathbb{E}\left(\sum_{i=1}^{n} w_n(x, X_i)\right)\mathbb{E}\left(\sup_{\theta\in\Theta} K(\theta, U)\right)\mathbb{E}\left(P\right)}{M_\varepsilon}\\ &= \frac{\mathbb{E}\left(\sup_{\theta\in\Theta} K(\theta, U)\right)(1-\tau_0)}{M_\varepsilon} \le \frac{M^*(1-\tau_0)}{M_\varepsilon} < \varepsilon.\end{aligned}$$

□

Lemma 8. *Under the assumptions of Theorem 1, it holds that*

$$\sup\left\{\|X_i - x\|_2 : w_n(x, X_i) > 0, i = 1, \dots, n\right\} = o_P(1).$$

Proof. This result follows from Lemma 2 of Wager and Athey (2018) which states that $\text{diam}(L_b(x)) = o_P(1)$. It does not require the random forest to be honest; i.e., we can assume that we use the same observations to place the splits and make predictions. For each tree $b = 1, \dots, B$ of the forest, we subsample $\mathcal{S}_b \subset \{1, \dots, n\}$ observations from the training data, with $|\mathcal{S}_b| = s < n$. Denote by $L_b(x) \subset \mathcal{X}$ the leaf containing the fixed predictor value $x \in \mathcal{X}$. Define the diameter $\text{diam}(L_b(x)) := \sup_{z,y\in L_b(x)}\|z - y\|_2$ of the leaf $L_b(x)$ as the length of the longest segment contained inside $L_b(x)$. Recall that the weights of a (not necessarily honest) random forest are defined as

$$w_n(x, X_i) = \frac{1}{B}\sum_{i=1}^{B} w_{n,b}(x, X_i) = \frac{1}{B}\sum_{b=1}^{B}\frac{\mathbb{1}\{X_i \in L_b(x), i \in \mathcal{S}_b\}}{|\{X_i \in L_b(x), i \in \mathcal{S}_b\}|}.$$

Also, note that

$$\left\{\|X_i - x\|_2 : w_n(x, X_i) > 0, i = 1, \ldots, n\right\} = \left\{\|X_i - x\|_2 : \exists b = 1, \ldots, B, X_i \in L_b(x), i \in \mathcal{S}_b\right\}$$
$$= \cup_{b=1}^{B} \left\{\|X_i - x\|_2 : X_i \in L_b(x), i \in \mathcal{S}_b\right\} \subset \cup_{b=1}^{B} \left\{\left\|y - x\right\|_2 : y \in L_b(x)\right\}.$$

Therefore,

$$\sup\left\{\|X_i - x\|_2 : w_n(x, X_i) > 0, i = 1, \ldots, n\right\} \leq \sup \cup_{b=1}^{B} \left\{\left\|y - x\right\|_2 : y \in L_b(x)\right\}$$
$$= \max_{b=1}^{B} \sup\left\{\left\|y - x\right\|_2 : y \in L_b(x)\right\} \leq \max_{b=1}^{B} \operatorname{diam}(L_b(x)).$$

Thus, for every $\varepsilon > 0$

$$0 \leq \mathbb{P}\left(\sup\left\{\|X_i - x\|_2 : w_n(x, X_i) > 0, i = 1, \ldots, n\right\} > \varepsilon\right)$$
$$\leq \mathbb{P}\left(\max_{b=1}^{B} \operatorname{diam}(L_b(x)) > \varepsilon\right) \leq \sum_{b=1}^{B} \mathbb{P}\left(\operatorname{diam}(L_b(x)) > \varepsilon\right) \to 0.$$

□

B Partial Derivative on the Boundary

For any function $f : [0,1]^p \to \mathbb{R}$, we define the first order partial derivative on the boundary by

$$\partial_{x_j} f(x) := \begin{cases} \lim_{h \downarrow 0} \frac{f(x + h e_j)}{h}, & \text{if } x \in [0,1]^p, x_j = 0, \\ \lim_{h \downarrow 0} \frac{f(x) - f(x - h e_j)}{h}, & \text{if } x \in [0,1]^p, x_j = 1. \end{cases}$$

C Weight Function Estimation

In quantile regression tasks, the weight function $(x, y) \mapsto w_n(x, y)$ estimated by GRF measures the similarity between x and y according to their conditional distribution.

Figure 9 shows the localizing weights $w_n(x, X_i)$, $x, X_i \in \mathbb{R}^p$, for two test predictors x with $x_1 = -0.2, 0.5$, respectively. The data is generated according to Example 1, with $n = 2000$ observations and $p = 40$ predictors. In the left panel of Figure 9, the observations (X_i, Y_i) with $X_{i1} < 0$ are the ones influencing most the test predictor x with $x_1 = -0.2$. This is because they share the same conditional distribution. A similar argument holds for the right panel of Figure 9.

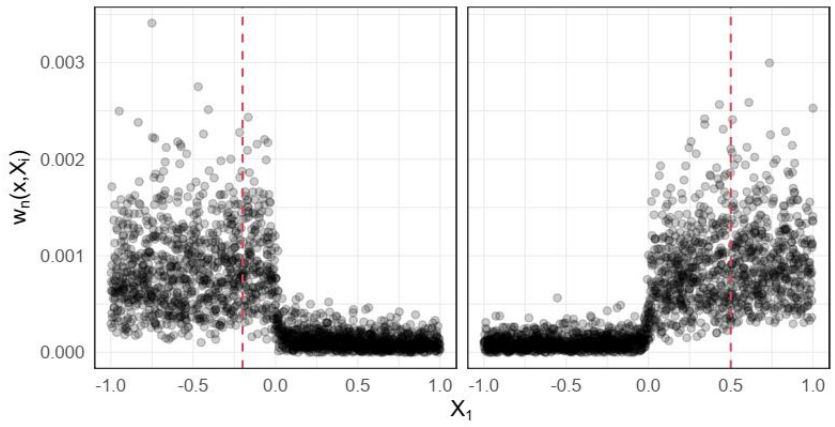

Figure 9: The height of the points represents the localizing weights $w_n(x, X_i)$ between a test predictor $x \in \mathbb{R}^p$ and each training observation $X_i \in \mathbb{R}^p$. The dashed line indicates the first coordinate of the test predictor values.

D Additional Material for Simulation Study

D.1 Sensitivity of Intermediate Threshold Level

Figure 10 shows the square root MISE of predicted quantiles as a function of the intermediate threshold τ_0 for different quantile levels τ and different shape parameters ξ of the noise variable. Even though the threshold choice has an influence on the prediction accuracy, from the scales of the square root MISE it can be seen that this influence is not too strong. The optimal choice will depend on the properties of the data such as the tail heaviness of the response; for details see de Haan and Ferreira (2006, Section 3.2). In applications, there are numerous data-driven methods for choosing the threshold such as the mean excess plot (see Embrechts et al., 2012, Section 6.2.2).

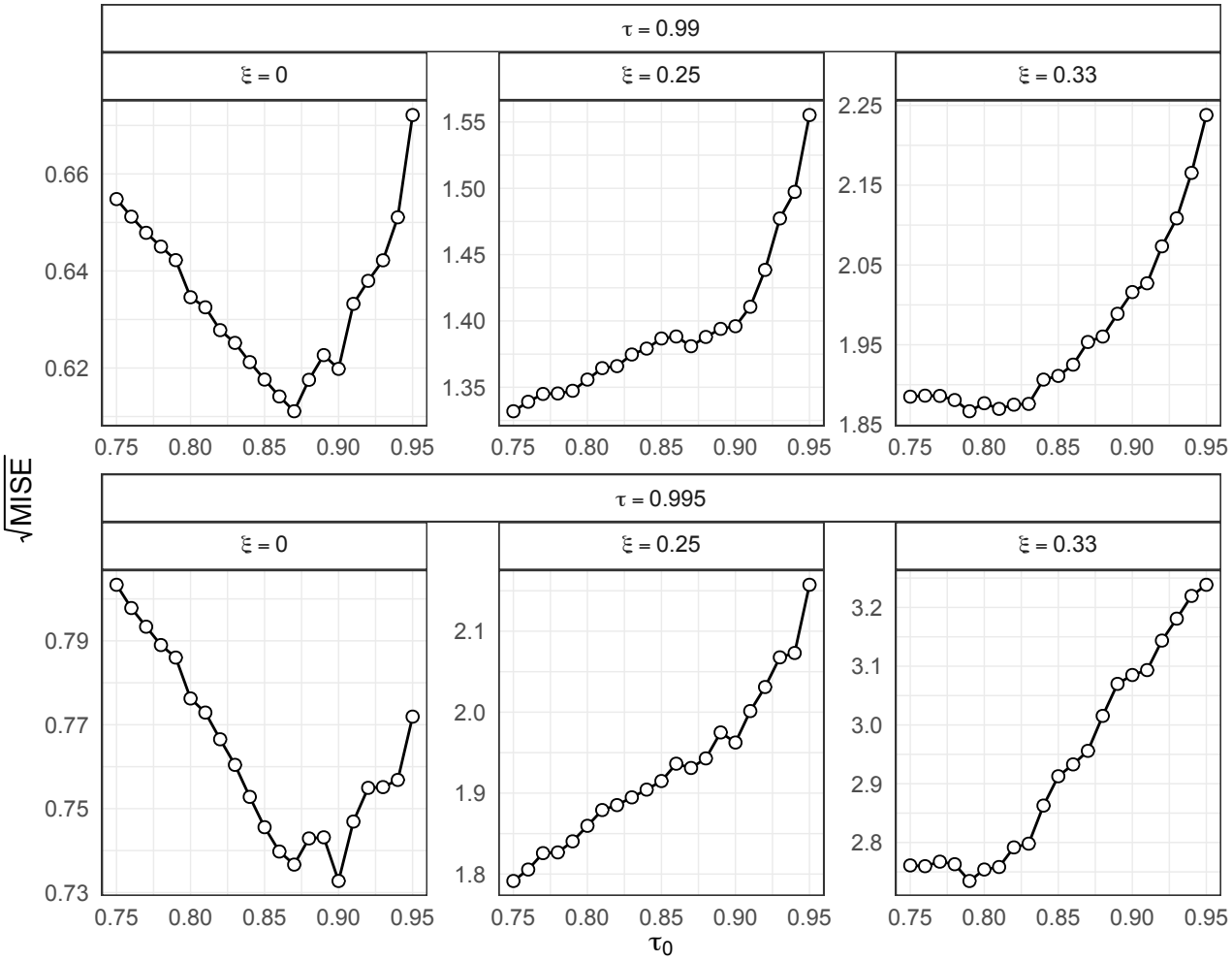

Figure 10: Square root MISE of predicted quantiles as a function of the intermediate threshold τ_0 for different quantile levels τ and different shape parameters ξ of the noise variable. Each point is an average over $m = 100$ repetitions. The data is generated according to Example 1 in the main text, where we set the dimension of the predictor space to $p = 5$.

D.2 Experiment 3

In the last experiment mentioned in Section 4, we consider more complex regression functions depending on more signal variables both in the scale and shape parameters. While the predictor variables X are uniform distributed on $[-1, 1]^p$ with $p = 10$, the conditional response follows three different models

$$(Y \mid X = x) \sim s_j(x) T_{\nu(x)}, \quad j = 1, 2, 3,$$

where we allow both degrees of freedom $\nu(x)$ and the scale $s_j(x)$ of the Student's t distribution to depend on the predictors. In particular, we model the degrees of freedom as a decreasing function

of the first predictor as $\nu(x) = 3[2 + \tanh(-2x_1)]$, and the different scale functions as

$$
\begin{aligned}
s_1(x) &= [2 + \tanh(2x_1)](1 + x_2/2), \\
s_2(x) &= 4 - (x_1^2 + 2x_2^2), \\
s_3(x) &= 1 + 2\pi\varphi(2x_1, 2x_2),
\end{aligned}
$$

where φ denotes a centered bivariate Gaussian density with unit variance and correlation coefficient equal to 0.75. The first scale function $s_1(x)$ is non-linear with respect to the first predictor and contains an interaction effect between the first two predictors. The function $s_2(x)$ is quadratic and decreasing in the first two dimensions. The third scale function $s_3(x)$ is non-linear in the first two predictors and contains an interaction effect. The sample size is $n = 5000$.

In this experiment we compare ERF, GRF, GBEX, EGP Tail and the unconditional method. We leave out EGAM because we observed it performs poorly in the scenarios considered here. Figure 11 shows the boxplots of $\sqrt{\text{ISE}}$ over $m = 50$ simulations over different models, methods, and quantile levels. For better visualization, we remove large outliers of GRF, QRF, and EGP Tail. We observe that ERF and GBEX generally outperform the other methods over all models and quantile levels, where GBEX has a slight advantage in high quantiles for Models 2 and 3. GRF and QRF seems to deteriorate completely for very large quantiles.

E Additional Material for U.S. Wage Analysis

E.1 Additional Figure

Figure 12 shows that estimated GPD parameters $\hat{\theta}(x)$ for the original response as a function of age for groups with less or more than 15 years of education.

E.2 Analysis with Log-Transformed Response

Following Angrist et al. (2009), we consider here the natural logarithm of the wage as response variable for quantile regression. We perform the same analysis as in Section 5 again with this log-transformed response since it highlights several interesting properties of the ERF algorithm. Figure 13 shows the GPD parameters $\hat{\theta}^{\log}(x)$ estimated by ERF as a function of years of education when the response is $\log(Y)$. We notice that the log-transformation makes the response lighter-tailed, with estimated shape parameters $\hat{\xi}^{\log}(x)$ fairly close to 0. The scale parameters $\hat{\sigma}^{\log}(x)$ still show a certain structure, but they vary on a much smaller scale compared to $\hat{\sigma}(x)$ estimated on the original response; see Figure 6 in the main text. These observations are consistent with theory since it is well-known that the log-transformation renders heavy-tailed data into light-tailed (Embrechts et al., 2012, Example 3.3.33). Moreover, the shape parameter on the original data then essentially acts as a scale parameter in the GPD approximation of the log-transformed data, explaining the smaller variation of $\hat{\sigma}^{\log}(x)$.

Figure 14 in the main text shows the (exponentiated) predicted quantiles $\exp\{\hat{Q}_x^{\log}(\tau)\}$ of the different methods as a function of years of education when the response is $\log(Y)$; we removed

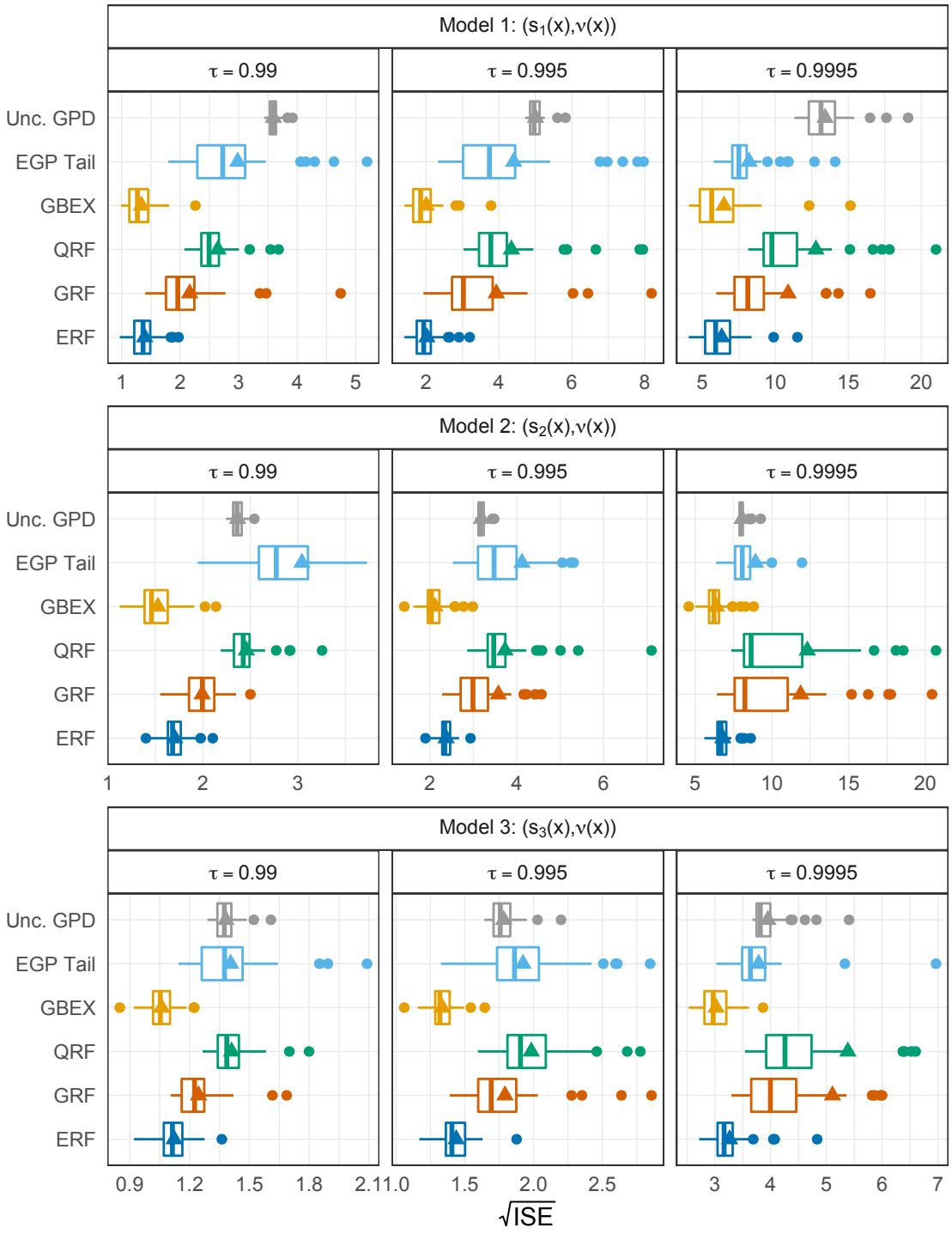

Figure 11: Boxplots of $\sqrt{\text{ISE}}$ over $m = 50$ simulations for different generative models (rows) and quantile levels (columns). The predictor space dimension is set to $p = 10$. Triangles represent the average values.

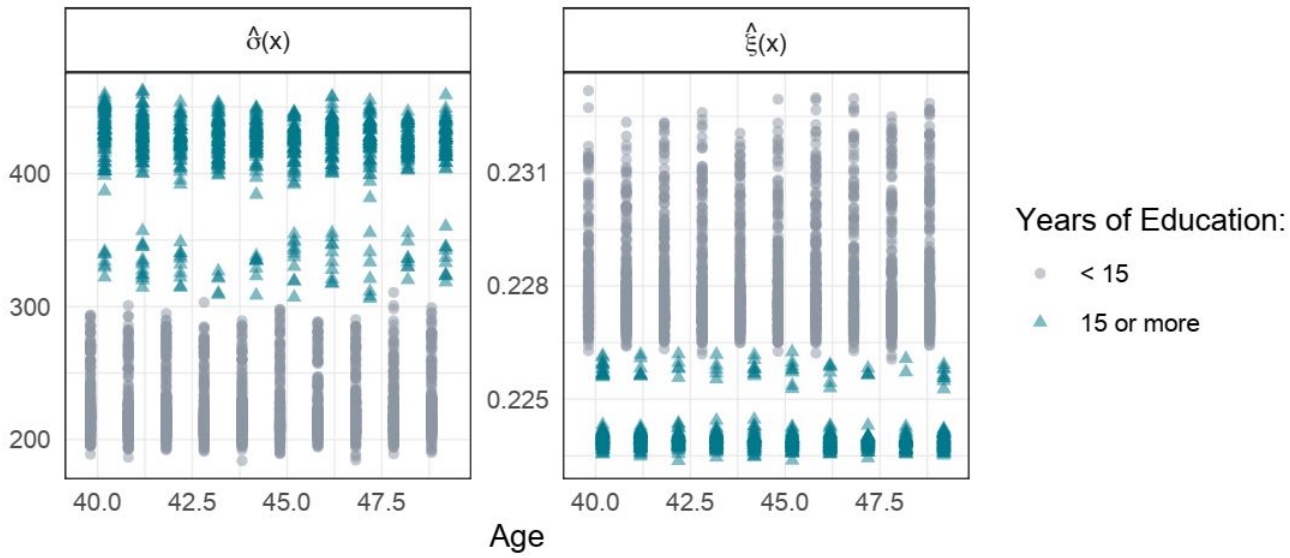

Figure 12: Estimated GPD parameters $\hat{\theta}(x)$ as a function of age for groups with less (circles) or more (triangles) than 15 years of education.

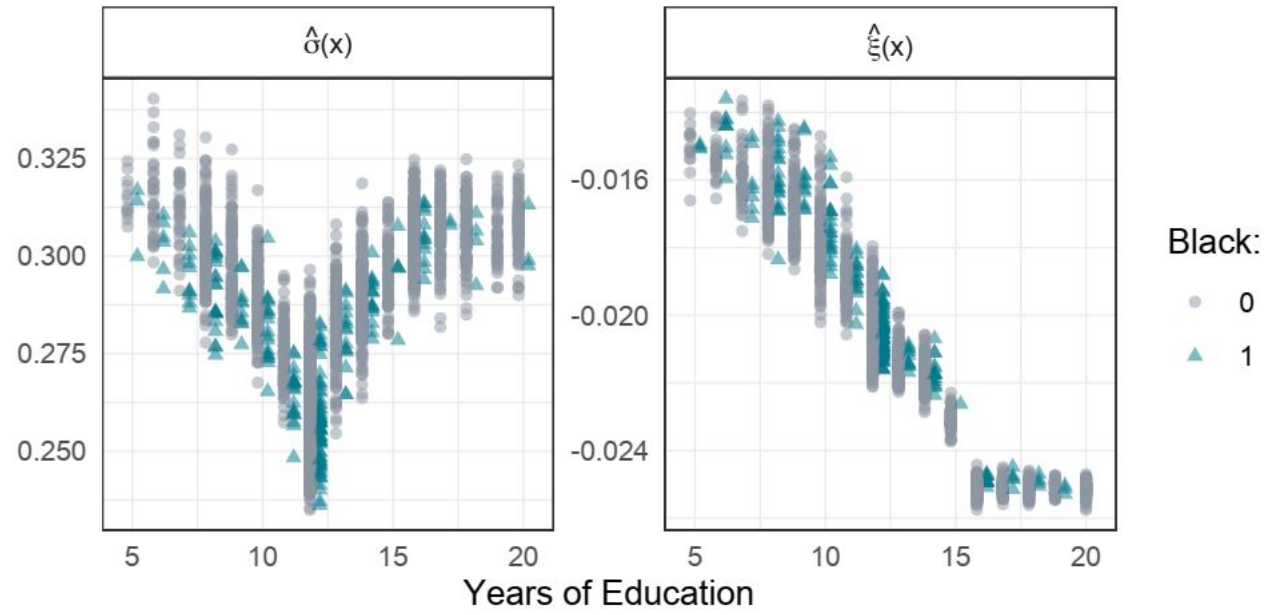

Figure 13: Estimated GPD parameters $\hat{\theta}(x)$ for the log-response as a function of the years of education for the black (triangles) and white (circles) subgroups.

again all quantiles above 6,000 predicted by GRF. By construction, GRF is invariant to the log-transformation, while the methods based on extrapolation may produce predictions that differ from $\hat{Q}_x(\tau)$ in Figure 7 fitted on the original data. The reason is that the approximation by the GPD is done on heavy-tailed data on the original scale and on much lighter-tailed data on the log-scale. We

observe in Figure 14 that the flexible methods ERF and GBEX have the desirable property that the predictions do not change much under marginal transformations. The unconditional method on the other hand seems to be sensitive to marginal transformation and works better on the log-transformed data as it captures a larger variability of the conditional quantiles even for high τ. This is confirmed by Figure 15 where we observe that the unconditional method has a smaller loss especially for higher quantiles, while all other methods have a similar performance as on the original data. To better understand this behavior, we recall the GPD approximation from (1.1) for large quantiles estimated on the original response as

$$\hat{Q}_x(\tau) \approx \hat{Q}_x(\tau_0) + G^{-1}\left(\frac{\tau - \tau_0}{1 - \tau_0}; \hat{\theta}(x)\right), \tag{E.1}$$

where G^{-1} is the inverse of the distribution function (2.2) of the GPD; see Figure 7 in the main text. On the other hand, first estimating the quantiles of the log-transformed data with a similar approximation and then exponentiating these estimates results in

$$\exp\{\hat{Q}_x^{\log}(\tau)\} \approx \hat{Q}_x(\tau_0) \exp\left\{G^{-1}\left(\frac{\tau - \tau_0}{1 - \tau_0}; \hat{\theta}^{\log}(x)\right)\right\}, \tag{E.2}$$

where $\hat{\theta}^{\log}(x)$ is the parameter vector of the GPD fitted for the response $\log(Y)$; see Figure 14. We note that $\hat{Q}_x(\tau_0)$ is the same in both approximations since it is fitted using quantile GRF, which is invariant under marginal transformations. Comparing (E.1) and (E.2) shows that the intermediate quantiles have an additive and multiplicative influence on the extreme quantiles, respectively. This explains why using the unconditional method for the GPD with $\hat{\theta}^{\log}(x) \equiv \hat{\theta}^{\log}$ seems to work better on the log-transformed data. Indeed, the different multiplicative scalings observed for ERF and GBEX in Figure 7 in the main text cannot be represented by (E.1) with unconditional GPD, but they can be represented by (E.2) if the intermediate quantile already carries the structure.

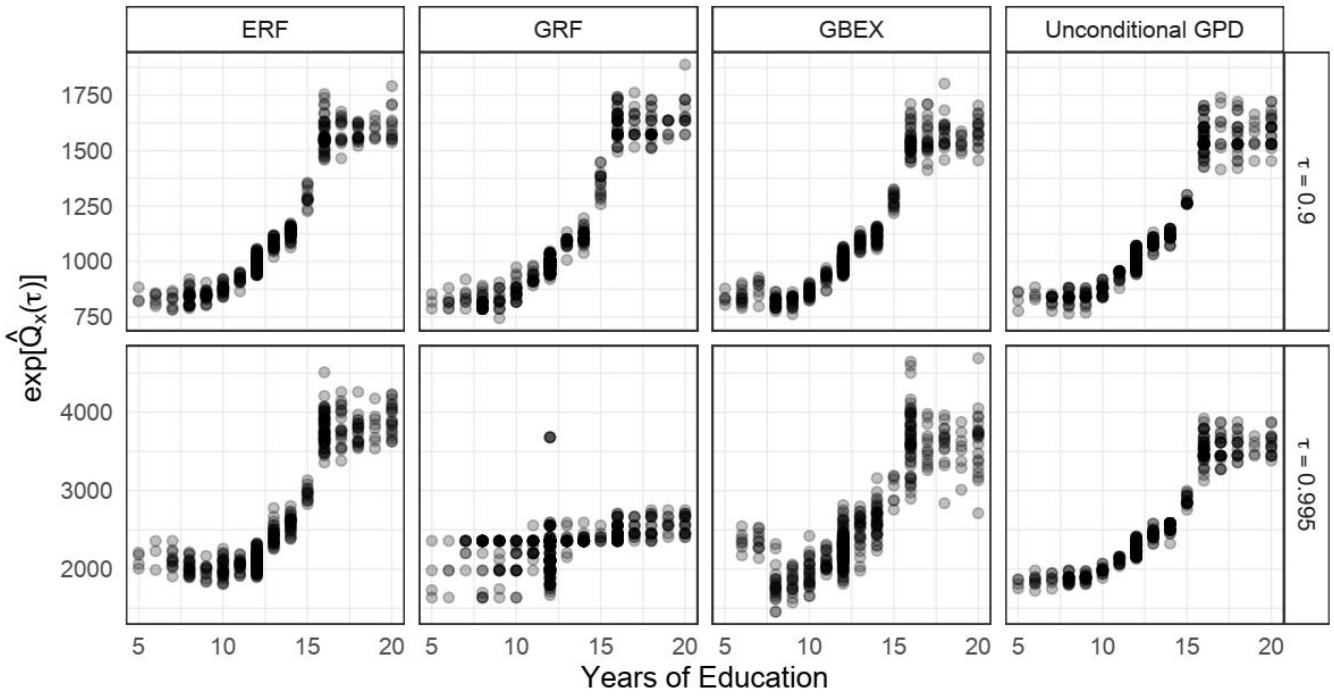

Figure 14: Predicted quantiles at levels $\tau = 0.9, 0.995$ for ERF, GRF, GBEX, and the unconditional method fitted on the log-response.

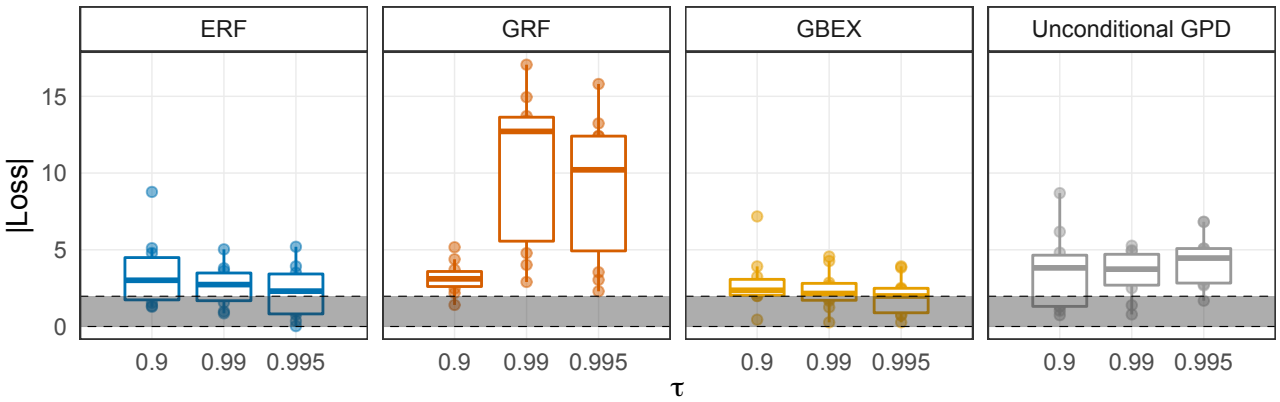

Figure 15: Absolute value of the loss (5.1) for the different methods fitted on the log-response of the U.S. wage data. The shaded area represents the 95% interval of the absolute value of a standard normal distribution.

References

J. D. Angrist, V. Chernozhukov, and I. Fernández-Val. Quantile regression under misspecification, with an application to the U.S. wage structure. *Econometrica*, 74(2):539–563, 2006. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/3598810.

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. URL https://doi.org/10.7910/DVN/JNEOLQ. https://doi.org/10.7910/DVN/JNEOLQ.

S. Athey, J. Tibshirani, and S. Wager. Generalized random forests. *The Annals of Statistics*, 47(2): 1148–1178, 2019. URL https://doi.org/10.1214/18-AOS1709.

A. A. Balkema and L. de Haan. Residual Life Time at Great Age. *The Annals of Probability*, 2 (5):792 – 804, 1974. doi: 10.1214/aop/1176996548. URL https://doi.org/10.1214/aop/1176996548.

J. Beirlant, T. D. Wet, and Y. Goegebeur. Nonparametric estimation of extreme conditional quantiles. *Statistical Computation and Simulation*, 74(8):567 – 580, 2004. doi: 10.1080/00949650310001623407. URL https://doi.org/10.1080/00949650310001623407.

J. Beirlant, G. Dierckx, and A. Guillou. Estimation of the extreme-value index and generalized quantile plots. *Bernoulli*, 11(6):949 – 970, 2005. doi: 10.3150/bj/1137421635. URL https://doi.org/10.3150/bj/1137421635.

G. Biau. Analysis of a random forests model. *Journal of Machine Learning Research*, 13(38): 1063–1095, 2012. URL http://jmlr.org/papers/v13/biau12a.html.

L. Breiman. Random forests. *Machine Learning, 45, 5–32*, 2001. ISSN 0885-6125. doi: 10.1023/A:1010933404324. URL http://dx.doi.org/10.1023/A:1010933404324.

A. Bücher and J. Segers. On the maximum likelihood estimator for the generalized extreme-value distribution. *Extremes*, 20(4):839–872, 2017. doi: 10.1007/s10687-017-0292-6. URL https://doi.org/10.1007/s10687-017-0292-6.

A. Bücher, J. Lilienthal, P. Kinsvater, and R. Fried. Penalized quasi-maximum likelihood estimation for extreme value models with application to flood frequency analysis. *Extremes*, pages 1–24, 2020. doi: 10.1007/s10687-020-00379-y. URL https://doi.org/10.1007/s10687-020-00379-y.

V. Chavez-Demoulin and A. C. Davison. Generalized additive modelling of sample extremes. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(1):207–222, 2005. doi: https://doi.org/10.1111/j.1467-9876.2005.00479.x. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9876.2005.00479.x.

V. Chernozhukov. Extremal quantile regression. *The Annals of Statistics*, 33(2):806 – 839, 2005. doi: 10.1214/009053604000001165. URL https://doi.org/10.1214/009053604000001165.

S. G. Coles and M. J. Dixon. Likelihood-based inference for extreme value models. *Extremes*, 2 (1):5–23, 1999.

A. Daouia, L. Gardes, S. Girard, and A. Lekina. Kernel estimators of extreme level curves. *Test, Spanish Society of Statistics and Operations Research/Springer*, 20(2):311 – 333, 2011. doi: 10.1007/s11749-010-0196-0.

L. de Haan and A. Ferreira. *Extreme Value Theory*. Springer, New York, 2006.

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

C. Dombry. Existence and consistency of the maximum likelihood estimators for the extreme value index within the block maxima framework. *Bernoulli*, 21(1):420 – 436, 2015. doi: 10.3150/13-BEJ573. URL https://doi.org/10.3150/13-BEJ573.

C. Dombry and A. Ferreira. Maximum likelihood estimators based on the block maxima method. *Bernoulli*, 25(3):1690–1723, 2019. ISSN 1350-7265. doi: 10.3150/18-BEJ1032. URL https://doi.org/10.3150/18-BEJ1032.

H. Drees, A. Ferreira, and L. de Haan. On maximum likelihood estimation of the extreme value index. *Ann. Appl. Probab.*, 14(3):1179–1201, 2004. ISSN 1050-5164. doi: 10.1214/105051604000000279. URL https://doi.org/10.1214/105051604000000279.

P. Embrechts, C. Klüppelberg, and T. Mikosch. *Modelling Extremal Events for Insurance and Finance*. Stochastic Modelling and Applied Probability. Springer Heidelberg New York Dordrecht London, 9^{th} edition, 2012. ISBN 978-3-540-60931-5. doi: 10.1007/978-3-642-33483-2.

S. Engelke, R. de Fondeville, and M. Oesting. Extremal behaviour of aggregated data with an application to downscaling. *Biometrika*, 106:127–144, 2019. doi: 10.1093/biomet/asy052.

S. Farkas, O. Lopez, and M. Thomas. Cyber claim analysis through generalized pareto regression trees with applications to insurance pricing and reserving. Preprint at https://hal.archives-ouvertes.fr/hal-02118080v2, 2020.

A. Ferreira, L. de Haan, and C. Zhou. Exceedance probability of the integral of a stochastic process. *J. Multivariate Anal.*, 105:241 – 257, 2012.

J. H. Friedman. Greedy function approximation: a gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001.

J. H. Friedman. Stochastic gradient boosting. *Computational Statistics and Data Analysis*, 38(4): 367–378, 2002.

L. Gardes and G. Stupfler. An integrated functional Weissman estimator for conditional extreme quantiles. *REVSTAT*, 17(1):109–144, 2019. ISSN 1645-6726. doi: 10.1007/s10687-013-0174-5. URL https://doi.org/10.1007/s10687-013-0174-5.

J. H. Halton. Algorithm 247: Radical-inverse quasi-random point sequence. *Commun. ACM*, 7(12):701–702, Dec. 1964. ISSN 0001-0782. doi: 10.1145/355588.365104. URL https://doi.org/10.1145/355588.365104.

T. J. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, New York, NY, USA, second edition, 2009.

P. J. Heagerty and M. S. Pepe. Semiparametric estimation of regression quantiles with application to standardizing weight for height and age in us children. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 48(4):533–551, 1999.

W. Hoeffding. A Class of Statistics with Asymptotically Normal Distribution. *The Annals of Mathematical Statistics*, 19(3):293 – 325, 1948. doi: 10.1214/aoms/1177730196. URL https://doi.org/10.1214/aoms/1177730196.

R. Koenker. Additive models for quantile regression: Model selection and confidence bandaids. *Brazilian Journal of Probability and Statistics*, 25(3):239 – 262, 2011. doi: 10.1214/10-BJPS131. URL https://doi.org/10.1214/10-BJPS131.

R. Koenker and G. Bassett. Regression quantiles. *Journal of the Econometric Society*, 46(1):33–50, 1978.

C. Martins-Filho, F. Yao, and M. Torero. High-order conditional quantile estimation based on nonparametric models of regression. *Econometric Reviews*, 34(6 - 10):907 – 958, 2015. doi: 10.1080/07474938.2014.956612.

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

W. K. Newey. Uniform convergence in probability and stochastic equicontinuity. *Econometrica*, 59 (4):1161–1167, 1991. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/2938179.

J. I. Pickands. Statistical inference using extreme value order statistics. *Annals of Statististics*, 1975.

E. Scornet, G. Biau, and J.-P. Vert. Consistency of random forests. *The Annals of Statistics*, 43(4):1716 – 1741, 2015. doi: 10.1214/15-AOS1321. URL https://doi.org/10.1214/15-AOS1321.

R. L. Smith. Maximum likelihood estimation in a class of nonregular cases. *Biometrika*, 72(1): 67–90, 1985. ISSN 00063444. URL http://www.jstor.org/stable/2336336.

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

C. J. Stone. Optimal Rates of Convergence for Nonparametric Estimators. *The Annals of Statistics*, 8(6):1348 – 1360, 1980. doi: 10.1214/aos/1176345206. URL https://doi.org/10.1214/aos/1176345206.

C. J. Stone. Optimal Global Rates of Convergence for Nonparametric Regression. *The Annals of Statistics*, 10(4):1040 – 1053, 1982. doi: 10.1214/aos/1176345969. URL https://doi.org/10.1214/aos/1176345969.

M. Taillardat, A.-L. Fougères, P. Naveau, and O. Mestre. Forest-based and semiparametric methods for the postprocessing of rainfall ensemble forecasting. *Weather and Forecasting*, 34(3):617 – 634, 2019. doi: 10.1175/WAF-D-18-0149.1. URL https://journals.ametsoc.org/view/journals/wefo/34/3/waf-d-18-0149_1.xml.

J. W. Taylor. A quantile regression approach to estimating the distribution of multiperiod returns. *The Journal of Derivatives*, 7(1):64–78, 1999. ISSN 1074-1240. doi: 10.3905/jod.1999.319106. URL https://jod.pm-research.com/content/7/1/64.

J. W. Taylor. A quantile regression neural network approach to estimating the conditional density of multiperiod returns. *Journal of Forecasting*, 19(4):299–311, 2000. doi: https://doi.org/10.1002/1099-131X(200007)19:4⟨299::AID-FOR775⟩3.0.CO;2-V. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/1099-131X%28200007%2919%3A4%3C299%3A%3AAID-FOR775%3E3.0.CO%3B2-V.

J. Tibshirani, S. Athey, E. Sverdrup, and S. Wager. *grf: Generalized Random Forests*, 2021. URL https://CRAN.R-project.org/package=grf. R package version 2.0.2.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998. doi: 10.1017/CBO9780511802256.

J. Velthoen, J.-J. Cai, G. Jongbloed, and M. Schmeits. Improving precipitation forecasts using extreme quantile regression. *Extremes*, 22(4):599–622, 2019.

J. Velthoen, C. Dombry, J.-J. Cai, and S. Engelke. Gradient boosting for extreme quantile regression. *arXiv preprint arXiv:2103.00808*, 2021.

S. Wager and S. Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018. doi: 10.1080/01621459.2017.1319839. URL https://doi.org/10.1080/01621459.2017.1319839.

H. Wang and C.-L. Tsai. Tail index regression. *Journal of the American Statistical Association*, 104(487):1233–1240, 2009. doi: 10.1198/jasa.2009.tm08458. URL https://doi.org/10.1198/jasa.2009.tm08458.

H. J. Wang and D. Li. Estimation of extreme conditional quantiles through power transformation. *American Statistical Association*, pages 1062 – 1074, 2013. doi: 10.1080/01621459.2013.820134. URL https://doi.org/10.1080/01621459.2013.820134.

H. J. Wang, D. Li, and X. He. Estimation of high conditional quantiles for heavy-tailed distributions. *American Statistical Association*, pages 1453 – 1464, 2012. doi: 10.1080/01621459.2012.716382. URL https://doi.org/10.1080/01621459.2012.716382.

S. Yang. Censored median regression using weighted empirical survival and hazard functions. *Journal of the American Statistical Association*, 94(445):137–145, 1999. doi: 10.1080/01621459.1999.10473830. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1999.10473830.

B. D. Youngman. Generalized additive models for exceedances of high thresholds with an application to return level estimation for u.s. wind gusts. *Journal of the American Statistical Association*, 114(528):1865–1879, 2019. doi: 10.1080/01621459.2018.1529596. URL https://doi.org/10.1080/01621459.2018.1529596.

K. Yu and M. C. Jones. Local linear quantile regression. *Journal of the American Statistical Association*, 93(441):228–237, 1998. doi: 10.1080/01621459.1998.10474104. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1998.10474104.

K. Yu, Z. Lu, and J. Stander. Quantile regression: Applications and current research areas. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 52(3):331–350, 2003. ISSN 00390526, 14679884. URL http://www.jstor.org/stable/4128208.