Heavy-tailed density regression using the blended generalised Pareto distribution and xSPQR

Jordan Richards[1]   Reetam Majumder[2]

[1]School of Mathematics, University of Edinburgh

[2]Department of Mathematical Sciences, University of Arkansas

Extrapolation — peaks-over-threshold

- **Pickands–de Haan–Balkema Theorem:**
  High threshold excesses $Y - u \mid Y > u$ may be approximated by the generalised Pareto (GP) distribution: if there exists a scaling function $a(u) > 0$ such that as $u \to y_F$ (upper endpoint)

$$\Pr\left(\frac{Y-u}{a(u)} > y \mid Y > u\right) \to \text{a non-degenerate distribution}$$

$$\text{then it must be } 1 - F_{\mathrm{GP}}(y|\sigma_u, \xi) := \begin{cases} (1 + \xi y/\sigma_u)_+^{-1/\xi}, & \xi \neq 0, \\ \exp(-y/\sigma_u), & \xi = 0, \end{cases},$$

  where $\sigma_u > 0$ and $\xi \in \mathbb{R}$.

- In practice, we model excesses directly as $(Y - u) \mid Y > u \sim \mathrm{GP}(\sigma_u, \xi)$ where u is some high pre-specified threshold.

Regression

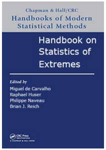
What if we have covariates $\mathbf{X} \in \mathbb{R}^p$?.

- Often make parametric assumptions about $Y|\mathbf{X} = \mathbf{x}$, e.g.,

$$(Y - u(\mathbf{x})) \mid (\mathbf{X} = \mathbf{x}, Y > u(\mathbf{x})) \sim \mathrm{GP}(\sigma_u(\mathbf{x}), \xi(\mathbf{x})),$$

  with $u(\mathbf{x}) > 0$ some varying threshold function.
- Lots of **AI-based** options:
  - **neural networks**, e.g., Allouche et al. (2024), Cisneros et al. (2024), Pasche and Engelke (2024), Richards and Huser (2025).
  - trees (Farkas et al., 2024) and forests (Gnecco et al., 2024)
  - boosting (Velthoen et al., 2023; Koh, 2023)
  - GAMs (Chavez-Demoulin and Davison, 2005; Youngman, 2019)

- Richards, J. and Huser, R. (2025). **Extreme Quantile Regression with Deep Learning.** In *Handbook on Statistics of Extremes*, Chapman & Hall/CRC.



Conditional setting

What if we have covariates $\mathbf{X} \in \mathbb{R}^p$?.

- Often make parametric assumptions about $Y|\mathbf{X} = \mathbf{x}$, e.g.,

$$(Y - u(\mathbf{x})) \mid (\mathbf{X} = \mathbf{x}, Y > u(\mathbf{x})) \sim \mathrm{GP}(\sigma_u(\mathbf{x}), \xi(\mathbf{x})),$$

  with $u(\mathbf{x}) > 0$ some varying threshold function.

- What about i) below the threshold, ii) choosing the threshold, iii) interpretability?

- We propose a semi-parametric density regression model that has GP upper-tails **without the need for threshold selection**.

Background – SPQR

Introduced by Xu and Reich (2021), SPQR is a flexible, semi-parametric approach to conditional density estimation.

- **No parametric assumptions**; instead, the conditional density is a convex combination of M-spline basis functions:

$$f_{\mathrm{SPQR}}(y|\mathbf{x}) = \sum_{k=1}^{K} w_k(\mathbf{x}) M_k(y),$$

with weights $w_k(\mathbf{x}) : \mathbb{R}^p \mapsto [0,1], k = 1, \ldots, K$, satisfying $\sum_{k=1}^{K} w_k(\mathbf{x}) = 1$ for all $\mathbf{x}$.

Xu, S.G. and Reich, B. J. (2021). Bayesian nonparametric quantile process regression and estimation of marginal quantile effects. *Biometrics*, 79:151–164

- Each basis function, $M_k(y)$, is a **valid PDF** on $[0,1]$ (Ramsay, 1988).
- The integral of an M-spline is an I-spline:

$$F_{\mathrm{SPQR}}(y|\mathbf{x}) = \sum_{k=1}^{K} w_k(\mathbf{x}) I_k(\mathbf{x}).$$

- The weights $\mathcal{W}(\mathbf{x}) := \{w_1(\mathbf{x}), \ldots, w_K(\mathbf{x})\}$ are modelled as a MLP with softmax final layer.
- Although very flexible, and fast-to-compute, F_{SPQR} satisfies no asymptotic guarantees and has bounded support.

Ramsay, J. O. (1988). Monotone regression splines in action. *Statistical Science*, 3(4):425–441

Blended Generalised Pareto

- Castro-Camilo et al. (2022) proposed the blended generalised extreme value distribution (bGEV), which blends the **Gumbel** and **Fréchet** distributions

  $\Rightarrow$ the resulting distribution function has an exact **Gumbel** lower-tail and **Fréchet** upper-tail.
- We follow a similar idea, but instead blend the GP distribution with a constituent *bulk* distribution, say F_{bulk}.
- Here we present the specific case of the **unconditional** blended GP with $F_{\text{bulk}} := F_{\text{SPQR}}$; we will introduce covariates later.

Play along at home!

You can also follow the link `https://reetamm-xspqr.share.connect.posit.cloud`

Blended Generalised Pareto

We define a bGP$(\mathcal{W},\xi)$ r.v. via its **continuous** distribution function

$$H(y|\mathcal{W},\xi) = \begin{cases} F_{\mathrm{SPQR}}(y|\mathcal{W})^{1-p(y)} F_{\mathrm{GP}}(y-\tilde{u}|\tilde{\sigma}_u,\xi)^{p(y)}, & y > \tilde{u}, \\ F_{\mathrm{SPQR}}(y|\mathcal{W}), & y \leq \tilde{u}, \end{cases} \tag{1}$$

where $p(y) \in [0,1]$ is a weighting function;

$$p(y) = p(y; a, b, c_1, c_2) = F_{\mathrm{Beta}}\left(\frac{y-a}{b-a}, c_1, c_2\right),$$

where $F_{\mathrm{Beta}}(\cdot, c_1, c_2)$ is a Beta(c_1, c_2) dist. with shapes $c_1 > 3, c_2 > 3$.

Note that $p(y) = 0$ for any $y < a$ and $p(y) = 1$ for any $y > b$.

Blended generalised Pareto

- We blend F_{SPQR} and F_{GP} in the interval $[a, b] \subset [0, 1]$, where the bounds are the p_a and p_b quantiles of F_{SPQR} $(p_b > p_a)$.
- To ensure continuity of H, we require

$$p_a := F_{\mathrm{SPQR}}(a|\mathcal{W}) = F_{\mathrm{GP}}(a - \tilde{u}|\tilde{\sigma}_u, \xi)$$
$$p_b := F_{\mathrm{SPQR}}(b|\mathcal{W}) = F_{\mathrm{GP}}(b - \tilde{u}|\tilde{\sigma}_u, \xi),$$

with:

$$(\tilde{\sigma}, \tilde{u}) = \begin{cases} \left(\frac{\xi(a-b)}{(1-p_a)^{-\xi}-(1-p_b)^{-\xi}}, a - \frac{(a-b)\{(1-p_a)^{-\xi}-1\}}{(1-p_a)^{-\xi}-(1-p_b)^{-\xi}}\right), & \xi \neq 0, \\ \left(\frac{(a-b)}{\log(1-p_a)-\log(1-p_b)}, a - \frac{(a-b)\{-\log(1-p_a)\}}{\log(1-p_a)-\log(1-p_b)}\right), & \xi = 0, \end{cases};$$

note that $\tilde{u} < a$.

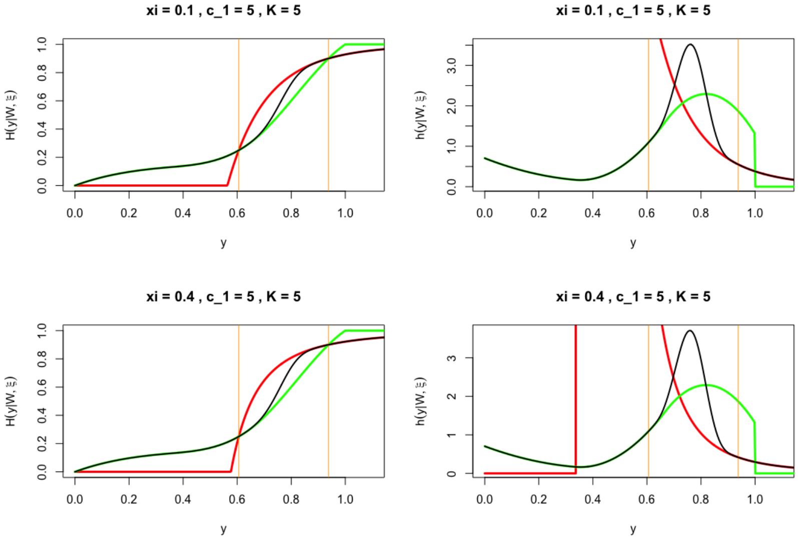
Blended GP

- For $\xi < 0$, the upper-endpoint of $H(\cdot|\mathcal{W},\xi)$ satisfies $\tilde{u} - \tilde{\sigma}_u/\xi > b$; for $\xi \geq 0$, the upper-endpoint of $H(\cdot|\mathcal{W},\xi)$ is infinite.

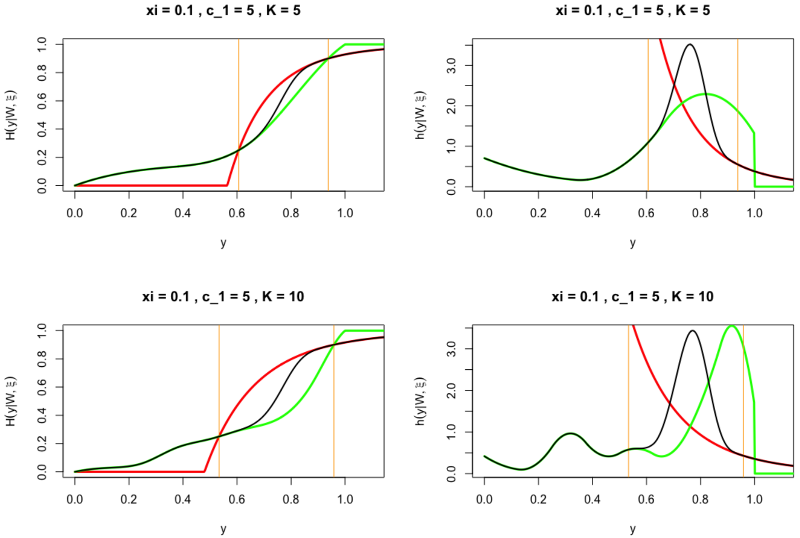
- The density is closed-form, and is **smooth** and **continuous**.

Increasing tail-heaviness

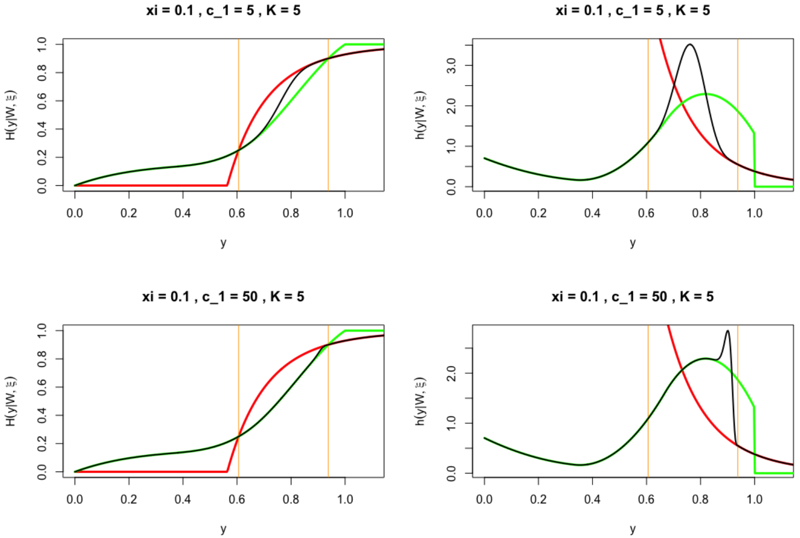
Left: bGP, GP, SPQR distribution. Right: corresponding density functions.



Increasing bulk-flexibility

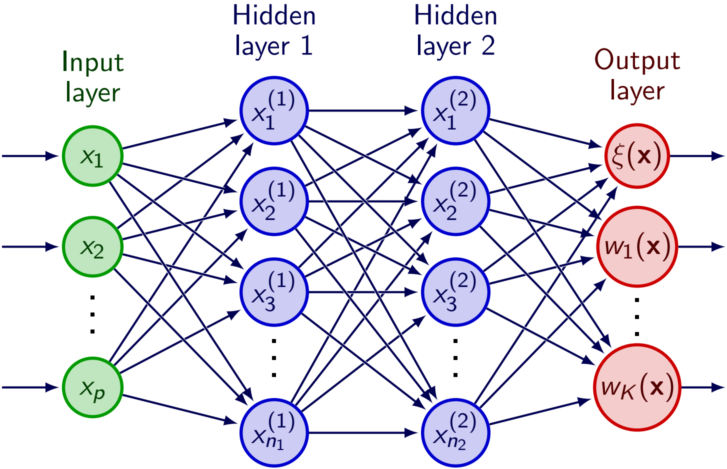


Increasing SPQR weighting



xSPQR

In the presence of covariates, we model $\mathbf{x} \mapsto (\xi(\mathbf{x}), \mathcal{W}(\mathbf{x}))$ via an MLP:



We refer to this framework as extremal SPQR (xSPQR).

Inference/coviarate importance

- Inference proceeds via maximum likelihood using Adam.
- xSPQR can be pre-trained with an SPQR fit.
- Via the `R` interface to `keras`.

- Variable importance (VI) can be the assessed for conditional quantile function $Q(\tau|\mathbf{x})$ at $\tau \in (0,1)$ **separately** of the shape $\xi(\mathbf{x})$.
- Using model-agnostic accumulated local effects (ALEs; Apley and Zhu, 2020).

Apley, D. W. and Zhu, J. (2020). Visualizing the effects of predictor variables in black box supervised learning models. *JRSSB*, 82:1059–1086

Simulation study

- Covariates $\mathbf{X}_i, i = 1, \ldots, 3$, are independent Unif(0,1).
- Response $Y \mid (\mathbf{X} = \mathbf{x})$ is log-normal$(\mu(\mathbf{x}), \sigma(\mathbf{x}))$ with

$$\mu(\mathbf{x}) = 5(1 - 1/[1 + \exp\{-(1 - 5x_1x_2)\}])$$

and

$$\sigma(\mathbf{x}) = 1/[1 + \exp\{-(1 - 5x_1x_2)\}].$$

- Only X_1 and X_2 act on Y.

- We take the MLP to have two layers, with n_h nodes and sigmoid activation in each layer.

Simulation study

- To evaluate estimation accuracy, we compute the integrated conditional 1-Wasserstein distance (IWD)

$$\text{IWD} = \int_{\mathcal{X}} \int_0^1 |Q(y|\mathbf{x}) - \hat{Q}(y|\mathbf{x})| \mathrm{d}\mathbf{x},$$

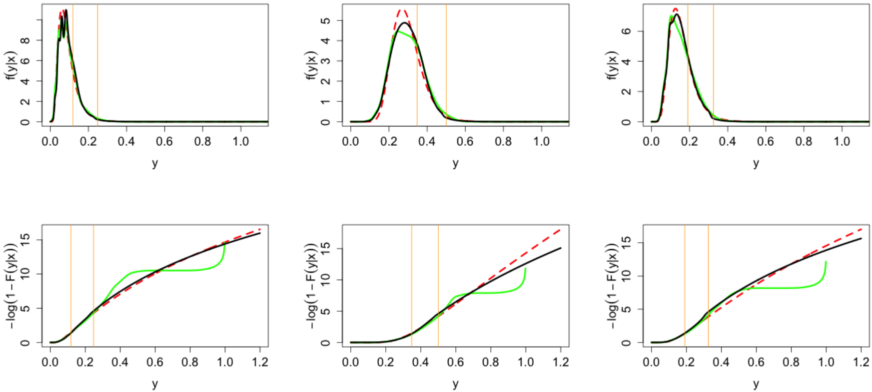
  where $\mathcal{X}$ is the sample space for $\mathbf{X}$ and $Q(y|\mathbf{x})$ denotes the conditional quantile function.
- We also consider a tail-calibrated version of the IWD, denoted by tIWD, which is constructed by replacing the limits of the inner integral of (17) with $[0.999, 1]$.

Results

n	K	n_h	tIWD	(p_a, p_b, c_1)
1000	15	16	11.2 (10.3, 12.3)/**9.23 (7.99, 10.6)**	$(0.9, 0.999, 5)$
	15	32	9.50 (8.63, 10.6)/**9.66 (8.18, 11.2)**	$(0.925, 0.999, 5)$
	25	16	12.0 (11.2, 13.0)/**9.56 (8.42, 10.9)**	$(0.925, 0.999, 5)$
	25	32	9.20 (8.31, 9.96)/**9.80 (8.70, 11.1)**	$(0.925, 0.999, 5)$
10000	15	16	10.6 (9.51, 11.3)/**7.08 (6.40, 7.85)**	$(0.75, 0.99, 25)$
	15	32	10.7 (10.0, 11.6)/**6.99 6.36, 8.05)**	$(0.75, 0.99, 25)$
	25	16	8.60 (7.33, 10.0)/**5.56 (4.45, 6.86)**	$(0.75, 0.99, 25)$
	25	32	10.2 (9.40, 16.6)/**5.29 (4.59, 6.50)**	$(0.75, 0.99, 25)$

Median (25%,75% quantiles) of tIWD estimates are reported for the original/**heavy-tailed** SPQR model, with the hyper-parameters (p_a, p_b, c_1) optimised for each row. Lower values are better.

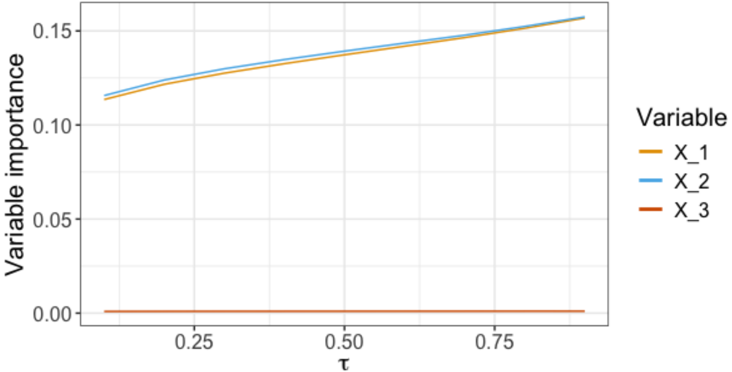
Test density estimation



Density (top) and log-survival (bottom) functions.

True, **SPQR**, and **xSPQR**.

Variable importance



VI scores (10^{-2}) for $\xi(\mathbf{x})$: $1.56, 2.34, 0.09$.

Case study: US wildfire burnt areas

- Burnt areas for over 10,000 moderate and large wildfires in the US, 1990–2020 (Lawler and Shaby, 2024).
- First and last 5 years used for testing. Model trained for 1995–2015.
- This leaves **6416 fires** for training and **3344 fires** for testing.

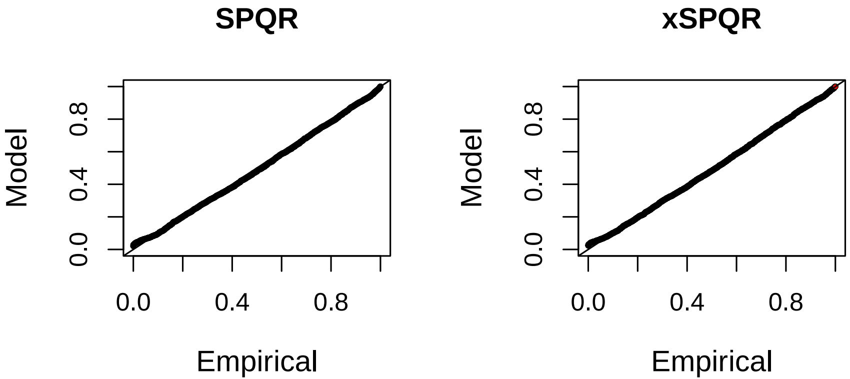
Lawler, E. S. and Shaby, B. A. (2024). Anthropogenic and meteorological effects on the counts and sizes of moderate and extreme wildfires. *Environmetrics*, 35(7):e2873

Case study: US wildfire burnt areas

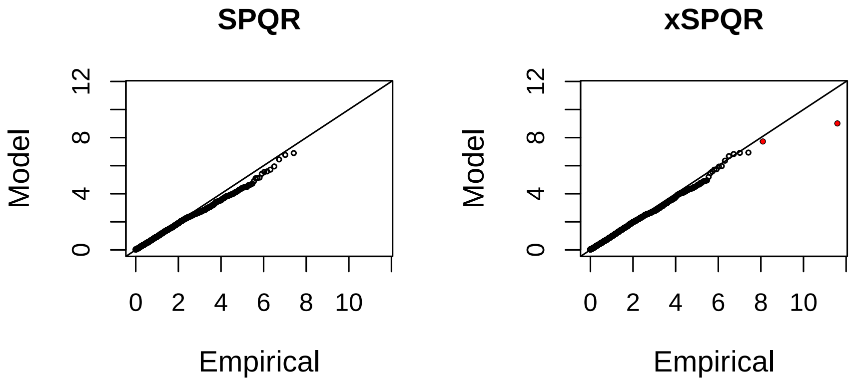
- We model the impacts of $\mathbf{X} =$
  - `pr_prev`: total precip. last year;
  - `pr_curr`: total precip. this month;
  - `rmin`: relative humidity;
  - `tmax`: maximum temperature;
  - `wspd`: windspeed;
  - `fire_yr`: fire year;

  on $Y = \sqrt{\text{Burnt area}}$.
- Model hyper-parameters/MLP architecture optimised via grid-search:
  - We here use a 2-layered MLP with $N_h = 12$ nodes per layer, sigmoid activations, and $K = 25$ basis functions.
  - We also constrain $\xi(\mathbf{x}) > 0$.

Model fits - bulk

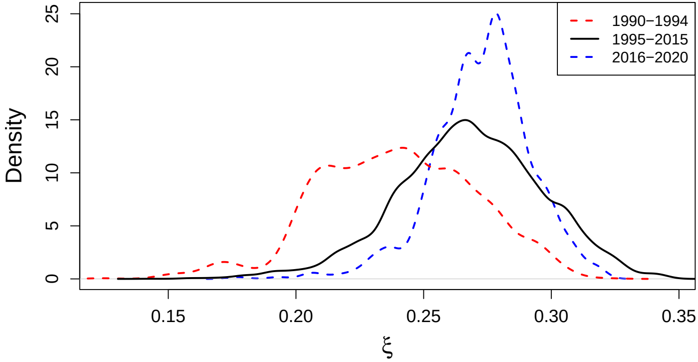


Model fits - tail



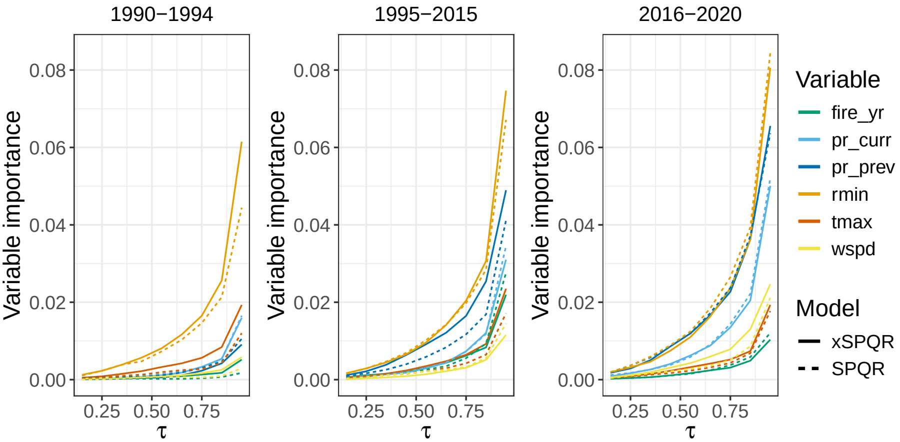
Red test points are impossible with SPQR.

Estimates of $\xi(\mathbf{x})$



Density of the estimated $\xi(\mathbf{x})$, stratified by time period.

Relative variable importance - bulk



Relative variable importance - tail

Time period	pr_prev	rmin	tmax	wspd	pr_curr	fire_yr
1990–1994	2.35	2.08	1.41	0.90	0.91	0.22
1995–2015	2.87	2.13	1.82	1.72	1.21	0.61
2016–2020	2.39	1.69	1.42	2.26	1.58	0.50

Spatial variation in quantiles

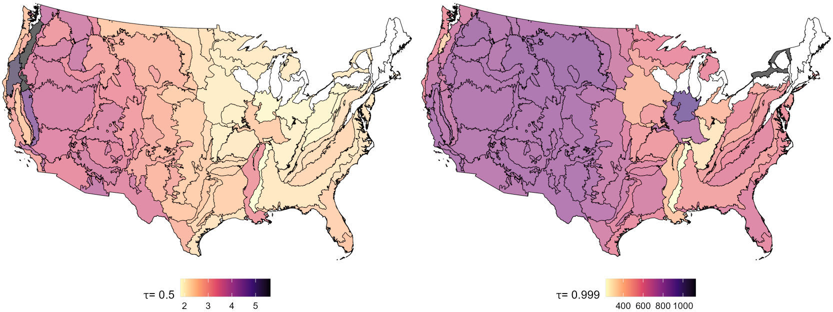


Figure: Estimates of the median (left) and 0.999-quantile (right) of burnt area (in 1000s of acres) for all observed wildfires, averaged over L3 ecoregions. Transparent regions do not include any observed wildfires.

U.S. Environmental Protection Agency. https://www.epa.gov/eco-research/level-iii-and-iv-ecoregions-continental-united-states

Conclusion

- Very flexible density regression model that is **EVT-compliant**.
- Requires no modelling of an intermediate exceedance threshold and provides a characterisation of the full density.
- Easily extendable to full real support and lower-tailed GP.
- Majumder, R. and Richards, J. (2025+). Semi-parametric bulk and tail regression using spline-based neural networks. arxiv:2504.19994.

SCAN ME

Special issue

Home > Collection

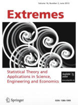
Bridging Heavy Tails and Artificial Intelligence

Participating journal: Extremes

Closed for submissions

The increasing frequency and severity of extreme events—such as catastrophic flooding, record-breaking temperatures, and unprecedented heatwaves—has highlighted the urgent need for innovative approaches in risk assessment and modeling. Modern advancements in data-collection techniques have provided increasingly large and complex datasets, which can only be processed used fast and scalable algorithms and computational software. This special issue aims to bridge the gap between Artificial Intelligence (AI) and Extreme Value Theory (EVT) to harness the strengths of both fields and address the growing challenges posed by these extreme events.

Participating journal



Journal

Extremes

Extremes is a dedicated platform for the publication of original research in statistical extreme value theory and its applications across various fields.

Publishing model	Hybrid
Journal Impact Factor	1.1 (2023)
Downloads	58k (2024)
Submission to first decision (median)	7 days

https://link.springer.com/collections/haghbdfdhb

References I

Allouche, M., Girard, S., and Gobet, E. (2024). Estimation of extreme quantiles from heavy-tailed distributions with neural networks. *Statistics and Computing*, 34(1):12.

Apley, D. W. and Zhu, J. (2020). Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society: Series B (Methodology)*, 82:1059–1086.

Castro-Camilo, D., Huser, R., and Rue, H. (2022). Practical strategies for generalized extreme value-based regression models for extremes. *Environmetrics*, 33(6):e2742.

Chavez-Demoulin, V. and Davison, A. C. (2005). Generalized additive modelling of sample extremes. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 54(1):207–222.

Cisneros, D., Richards, J., Dahal, A., Lombardo, L., and Huser, R. (2024). Deep graphical regression for jointly moderate and extreme Australian wildfires. *Spatial Statistics*, 59:100811.

Farkas, S., Heranval, A., Lopez, O., and Thomas, M. (2024). Generalized Pareto regression trees for extreme event analysis. *Extremes*, 27(3):437–477.

Gnecco, N., Terefe, E. M., and Engelke, S. (2024). Extremal random forests. *Journal of the American Statistical Association*, 119(548):3059–3072.

References II

Greenwell, B. M., Boehmke, B. C., and McCarthy, A. J. (2018). A simple and effective model-based variable importance measure. *arXiv preprint arXiv:1805.04755*.

Koh, J. (2023). Gradient boosting with extreme-value theory for wildfire prediction. *Extremes*, 26(2):273–299.

Lawler, E. S. and Shaby, B. A. (2024). Anthropogenic and meteorological effects on the counts and sizes of moderate and extreme wildfires. *Environmetrics*, 35(7):e2873.

Pasche, O. C. and Engelke, S. (2024). Neural networks for extreme quantile regression with an application to forecasting of flood risk. *The Annals of Applied Statistics*, 18(4):2818–2839.

Ramsay, J. O. (1988). Monotone Regression Splines in Action. *Statistical Science*, 3(4):425 – 441.

Velthoen, J., Dombry, C., Cai, J.-J., and Engelke, S. (2023). Gradient boosting for extreme quantile regression. *Extremes*, 26(4):639–667.

Xu, S. G. and Reich, B. J. (2021). Bayesian nonparametric quantile process regression and estimation of marginal quantile effects. *Biometrics*, 79:151–164.

Youngman, B. D. (2019). Generalized additive models for exceedances of high thresholds with an application to return level estimation for US wind gusts. *Journal of the American Statistical Association*, 114(528):1865–1879.

Appendix: construction of M-splines

Defined on a set of $K + d$ knots, $t_1, \dots, t_{K+d}$, which we take to be empirical quantiles of the training Y with equally-spaced levels.
For $d = 1$,

$$M_k(y|d) = \begin{cases} \frac{1}{t_{k+1} - t_k}, & t_k \leq y < t_{k+1}, \\ 0, & \text{otherwise.} \end{cases}$$

and, for $d > 1$,

$$M_k(y|d) = \frac{d[(y - t_k)M_k(y|d-1) + (t_{k+d} - y)M_{k+1}(y|d-1)]}{(d-1)(t_{k+d} - t_k)}.$$

For SPQR/xSPQR, $d = 3$.

Appendix: variable importance scores

Consider a generic differentiable function $g(\mathbf{x})$, where $\mathbf{x} = (x_1, \ldots, x_p)$ is the vector of covariates. The sensitivity of $g(\mathbf{x})$ to covariate x_j is quantified by the partial derivative

$$\dot{g}_j(x_j) = \frac{\partial g(\mathbf{x})}{\partial x_j}.$$

The accumulated local effect (ALE) of x_j on $g(\cdot)$ is then defined as

$$\mathrm{ALE}_j(x_j; g) = \int_{z_{0,j}}^{x_j} \mathbb{E}[\dot{g}_j(x_j) | x_j = z_j] \mathrm{d}z_j,$$

where $z_{0,j}$ is an approximate lower bound for x_j.

Appendix: variable importance scores (cont.)

Following Greenwell et al. (2018), we measure heterogeneity of the effect of X_j on $g(\cdot)$ by taking the standard deviation of $\mathrm{ALE}_j(X_j; g)$ with respect to X_j.

The variable importance (VI) score for X_j on $g(\cdot)$ is

$$\mathrm{VI}_j(g) = \sqrt{\mathrm{Var}_{X_j}[\mathrm{ALE}_j(X_j; g)]}.$$

For xSPQR, replace $g(\cdot)$ with the conditional τ-quantile function or $\xi(\mathbf{x})$.