Revista Colombiana de Estadística - Applied Statistics
January 2025, volume 48, issue 1, pp. 239–250
http://doi.org/10.15446/rce.v48n1.113610

Extreme Value Modeling with the Compound Poisson Process: Predicting Speeding Fine Collections

Modelado de valores extremos utilizando el proceso de poisson compuesto: predicción de la recaudación de multas por exceso de velocidad

Fernando Ferraz do Nascimento[a], Filipe Mateus S. Costa[b], Valmaria Rocha da Silva Ferraz[c]

Department of Statistics, Federal University of Piaui, Teresina, Brazil

Abstract

Speeding violations are intended as both punitive and educational actions for drivers who exceed the maximum allowed speed on the roads. From a tax collection perspective, they have a significant impact on the municipal budget. Extreme Value Theory has been a valuable tool for modeling the distribution of speeding violations. Additionally, it is equally important to model the daily number of speeding occurrences. A powerful method for jointly modeling both variables is the Compound Poisson Process. By understanding both speeding behavior and the number of infractions, we can estimate the expected value of total tax collections. A mixture of Gamma densities combining with the Generalized Pareto Distribution (GPD) in tail was proposed to model the distribution of speeding values. The results indicated significant potential for tax collection.

***Key words*:** Extreme value theory; Speeding violations; Compound Poisson process; Bayesian approach.

Resumen

Las infracciones por exceso de velocidad están destinadas tanto a acciones punitivas como educativas para los conductores que superan la velocidad máxima permitida en las carreteras. Desde una perspectiva de recaudación fiscal, tienen un impacto significativo en el presupuesto municipal.

[a]Ph.D. E-mail: fernandofn@ufpi.edu.br
[b]Bachelor's in Statistics. E-mail: filipemateus@outlook.com
[c]Ph.D. E-mail: valmaria@ufpi.edu.br

La Teoría de Valores Extremos ha sido una herramienta valiosa para modelar la distribución de las infracciones por exceso de velocidad. Además, es igualmente importante modelar el número diario de estas infracciones. Un método poderoso para modelar ambas variables conjuntamente es el Proceso de Poisson Compuesto. Al comprender tanto el comportamiento de los conductores como el número de infracciones, podemos estimar el valor esperado de la recaudación total de impuestos. Se propuso una mezcla de densidades Gamma combinada con la Distribución Generalizada de Pareto (GPD, por sus siglas en inglés) en la cola para modelar la distribución de los valores de exceso de velocidad. Los resultados indicaron un potencial significativo para la recaudación fiscal.

***Palabras clave*:** Enfoque bayesiano; Infracciones por exceso de velocidad; Proceso de Poisson compuesto; Teoría de valores extremos.

1. Introduction

The enforcement of traffic fines serves as a measure to deter behaviors that can disrupt the smooth flow of traffic, particularly in large cities. As a result of various discussions aimed at improving road organization, the Brazilian Traffic Code (BCT) regulates a comprehensive set of traffic rules. Speeding regulations, for instance, limit drivers from traveling at excessively high speeds, reducing the likelihood of accidents. According to Brazilian government agencies, there were 2,961,624 speeding violations in 2023 alone. In the same year, there were 65,176 traffic accidents, resulting in a total of 33,743 deaths, figures comparable to those of a country at war.

To mitigate this problem, a measure adopted by governments worldwide is to impose a maximum speed limit on major city roads. However, since many people do not adhere to this limit despite being aware of it, government agencies install speed cameras on the busiest streets, recording which cars exceed the limit and applying fines to these offenders. It is understood that if a person is not conscious of respecting the laws and staying within the limits, a financial penalty serves as an educational measure that works in most situations.

As a consequence of fines imposed for traffic violations, public revenue increases. For each traffic infraction, there is a specific fine amount, which escalates with the severity of the violation. In the case of speeding, the greater the excess speed, the higher the fine to be paid.

In all major Brazilian cities, speed cameras are positioned at various points along the main avenues. If a vehicle exceeds the permitted speed limit, the camera captures an image of the car's license plate, generating an infraction code and a corresponding fine amount to be paid by the vehicle owner.

In this work, we are interested in analyzing the behavior of two variables: the daily number of speeding violations and the distribution of these violations. To jointly analyze these two variables, the Compound Poisson Process will be applied in conjunction with Extreme Value Theory (EVT). Here, we will model the variable

N, representing the number of daily infractions, using a Poisson distribution, while the variable X of the speeding will be fitted using an extreme value model.

In this way, we aim to predict the average number of speeding violations that may occur per day, assess the probability of encountering a very high speeding violation, and, by combining these two variables, calculate the potential revenue that these speed cameras can generate for the municipality.

1.1. Extreme Value Theory

Extreme events are situations or behaviors that are rare but can have a significant impact on society when they do occur. They are the events that lie close to the tails of the distribution. Given this context, it is important that we can predict their occurrences. Studies on extreme values have been ongoing for nearly a century, beginning with the work of Fisher & Tippett (1928). von Mises (1954) and Jenkinson (1955) proposed the Generalized Extreme Value (GEV) distribution for modeling maxima.

The objective of Extreme Value Theory (EVT) is to analyze observed extreme values and estimate the probability of the occurrence of these events. Phenomena in which the probability of an extreme value is relatively high are characterized by distributions with heavy tails.

The Generalized Pareto Distribution (GPD), introduced by Pickands (1975), analyzes the distribution of excesses, considering a certain high threshold and has the following distribution function:

$$G(x|\xi,\eta,u) = \begin{cases} 1-\left(1+\xi\frac{(x-u)}{\eta}\right)^{-1/\xi}, & \text{if} \quad \xi \neq 0 \\ 1-\exp\big\{-(x-u)/\eta\big\}, & \text{if} \quad \xi = 0. \end{cases} \tag{1}$$

where $u > 0$, $\eta > 0$. The GPD is valid for $x > u$ for $\xi \geq 0$ and $u < x < (u - \eta/\xi)$ for $\xi < 0$. The parameters ξ, η and u represent shape, scale and location.

The GPD density is given by

$$g(x|\xi,\eta,u) = \begin{cases} \frac{1}{\eta}\left(1+\xi\frac{(x-u)}{\eta}\right)^{-(1+\xi)/\xi}, & \text{if} \quad \xi \neq 0 \\ \frac{1}{\eta}\exp\big\{-(x-u)/\eta\big\}, & \text{if} \quad \xi = 0. \end{cases} \tag{2}$$

The parameter ξ measures the weight of the tail, and the higher this value, the heavier the tail will be, resulting in a greater occurrence of extreme events. Figure 1 illustrates the tail weight for different values of ξ.

According to Coles (2001) and Embrechts et al. (1997), in EVT, it is crucial to find a method for determining the high quantiles above the threshold. If X follows a Generalized Pareto Distribution (GPD), it is important to know the probability of an event that is greater than or equal to q, i.e., $P(X > q) = 1 - p$. The p-quantile of the GPD distribution is given by:

$$q(p \mid \xi,\eta,u) = \begin{cases} u + \dfrac{((1-p)^{-\xi}-1)\eta}{\xi}, & \text{if} \quad \xi \neq 0 \\ u - \eta(\log(1-p)), & \text{if} \quad \xi = 0. \end{cases}$$

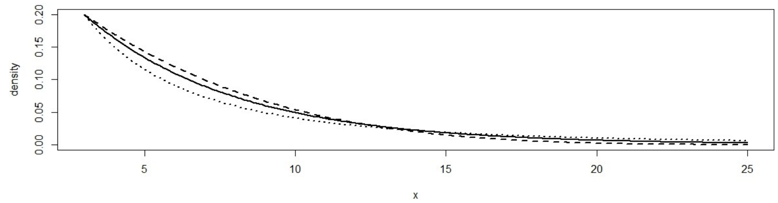

FIGURE 1: GPD density with $\xi = -0.3$ (dashed), $\xi = 0$ (full) and $\xi = 0.5$ (dotted).

1.2. The Compound Poisson Process

In the majority of problems in statistical inference, we consider a sample $X_1, \ldots, X_n$ that is independent and identically distributed (iid) with a F distribution, where the number of events n is fixed. However, in some situations, the number of occurrences in a given period can be considered random, making it necessary to estimate its distribution. Since this is a count variable, a natural approach is to use a Poisson distribution for N. Adelson (1966) is one of the first works to present the Compound Poisson Process.

The Compound Poisson Process is primarily used in insurance risk, where both the claim amount and the number of claims are random. An example of this is the work of Zhang et al. (2014). This distribution can also be applied in survival analysis, as shown by Ata & Ozel (2013). In this work, we will use the Compound Poisson Process in the context of extreme values, focusing on the daily number of speeding violations.

1.3. Speeding Violations

The data used in this work originate from measurements of excess speed recorded by radars in the city of Teresina, PI, Brazil, which is located in the northeastern region of the country and has a population of nearly 900,000 inhabitants. The radars whose data were analyzed are positioned on four busy avenues in the city, playing a vital role in reducing the number of accidents, as many drivers are discouraged from speeding due to the risk of receiving a fine.

Regarding the severity of the infraction, it occurs based on the amount of excess speed recorded; the greater the excess, the higher the fine to be paid. Based on this, the Brazilian Traffic Code (BCT) regulates three levels of severity according to the percentage of speed over the permitted limit registered during the vehicle's passage. Mild severity applies to measurements up to 20% above the allowed limit, high severity is for speeds between 20% and 50% over the limit, and very high severity is for speeds exceeding 50% of the limit. The increase in severity results in a higher fine amount. Tables 1 and 2 show the fines to be paid according to the severity level for roads with limits of 40 km/h and 60 km/h, respectively. Since we are primarily interested in the value of excess speed, we will work with the variable X representing the speeding value.

Table 1: Amount of the fine for speeding on roads with a limit of 40 km/h.

Type	%	Speeding	Fine
Medium	until 20%	$0 \leq x \leq 8$	130.16
Serious	between 20% and 50%	$8 \leq x \leq 20$	195.23
very Serious	higher than 50%	$x \geq 20$	880.41

Table 2: Amount of the fine for speeding on roads with a limit of 60 km/h.

Type	%	Speeding	Fine
Medium	until 20%	$0 \leq x \leq 12$	130.16
Serious	between 20% and 50%	$12 \leq x \leq 30$	195.23
very Serious	higher than 50%	$x \geq 30$	880.41

To calculate the probability of excess speed falling within each of these three ranges, we will first model all excesses x using Extreme Value Theory (EVT). With this distribution, we can calculate the probability of a vehicle exceeding the speed limit being in each interval of these tables. Thus, we can calculate the expected revenue from excess speed by multiplying the probability by the respective fine amount.

This work is organized as follows: Section 2 presents the Compound Poisson Process model applied to extreme values, along with the Bayesian inference procedure using MCMC. Section 3 presents the results of the proposed model as applied to speeding data. Section 4 summarizes the main conclusions of this work.

2. The Model

Consider that over several years, we have $N_1, \ldots, N_m$ representing the number of daily speeding violations at a given measurement point. On a specific day j, let $X_{j,1}, \ldots, X_{j,N_j}$ be the vector of speeding violations committed on that day. Therefore, for all days combined, we will have a vector

$$\mathbf{X} = (X_{1,1}, \ldots, X_{1,N_1}, X_{2,1}, \ldots, X_{2,N_2}, \ldots, X_{m,N_m})$$

which contains the total speeding violations recorded on all observed days.

The proposed model consists of performing the estimation in two stages. In the first stage, the daily number of excess speed violations $N = (N_1, \ldots, N_m)$ will be modeled using the Poisson distribution. In the second stage, the total vector of speed violations will be modeled by an extreme value distribution. We assume a separate modeling of these two variables because, given that the value of the excess has already occurred, the number of daily occurrences does not depend on the magnitude of the speeding violation. This same idea is applied in insurance data, where the number of claims is modeled independently of the magnitude of the claim.

2.1. The Compound Poisson Process

Initially, we will model the variable N, which refers to the daily number of infractions. The Poisson distribution is commonly used to model count data. According to Casella & Berger (2001), one of the basic assumptions upon which this model is developed is that, for small time intervals, the probability of an arrival is proportional to the waiting time.

Assuming that $N_j \sim Poisson(\lambda)$ and the vector of variables $\mathbf{N} = (N_1, N_2, N_3, \ldots, N_m)$ ais independent and identically distributed with a Poisson distribution, then the distribution of the sum

$$Y_j = \sum_{i=1}^{N_j} X_{i,j} \tag{3}$$

is known as the Compound Poisson Process, with mean and variance given by

$$E(Y) = E(N)E(X) \qquad VAR(Y) = E(N)Var(X) + (E(X))^2 Var(N) \tag{4}$$

In this work, a Bayesian framework was employed to obtain the posterior distribution of the parameter λ. Considering the likelihood function for the vector $\mathbf{N} = (N_1, N_2, N_3, \ldots, N_m)$, where each $N_j \sim Poisson(\lambda)$ and a prior distribution $\lambda \sim Gamma(a, b)$, the posterior distribution can be obtained as follows

$$\begin{aligned}
\pi(\lambda \mid N = n) &\propto \prod_{j=1}^{m} (P(N_j = n_j)) \times \lambda^{a-1} \exp(-b\lambda) \\
&\propto \prod_{j=1}^{m} (\lambda^{n_j} \exp(-\lambda)) \times \lambda^{a-1} \exp(-b\lambda) \\
&\propto \lambda^{\sum_{j=1}^{m} n_j} \exp(-m\lambda) \times \lambda^{a-1} \exp(-b\lambda) \\
&\propto \lambda^{\sum_{j=1}^{m} n_j + a} \exp(-(m+b)\lambda)
\end{aligned}$$

Thus, we see that the parameter λ has a posterior Gamma distribution, given by

$$\lambda \mid \mathbf{N} = n \sim Gamma(\sum_{j=1}^{m} n_j + a, m + b) \tag{5}$$

Considering that we have no prior information about λ, we will use a non-informative prior with $a = b$ set close to zero, resulting in a prior with high variance.

2.2. The $MPGD_k$ Model for Speeding

The second modeling step is to propose a distribution for the observed speeding violations. In Extreme Value Theory, values greater than a given threshold are estimated using the Generalized Pareto Distribution (GPD). For values below the threshold, several approaches can be proposed, one of which is the finite mixture of distributions.

Nascimento et al. (2012) proposed a model for extreme data that employs a non-parametric approximation based on a mixture of Gamma distributions for the non-tail. The finite mixture of Gammas, derived from the work of Wiper et al. (2001), is given by

$$h(x \mid \mu, \eta, p) = \sum_{j=1}^{k} p_j f_{\mathrm{G}}(x \mid \mu_j, \eta_j),$$

For the tail of the data, Nascimento et al. (2012) proposes a Generalized Pareto Distribution (GPD). Thus, the density of the model proposed by Nascimento et al. (2012), referred to as $MGPD_k$ is given by

$$f(x|\theta, p, \Psi) = \begin{cases} h(x|\mu, \eta, p), & if\ x\ \leq u \\ (1 - H(u|\mu, \eta, p))g(x|\Psi), & if\ x\ > u \end{cases} \tag{6}$$

where g is the density of the GPD distribution with parameters $\Psi = (\xi, \sigma, u)$, $\sigma > 0$. The GPD density is valid for $(x - u) \leq -\sigma/\xi$ and $x > u$.

The model proposed by Nascimento et al. (2012) has proven to be efficient in estimating the entire dataset and correctly identifying the true threshold using a Markov Chain Monte Carlo (MCMC) procedure.

In the database for this work, as we are dealing with speeding violations, it may seem intuitive to model all excesses using the Generalized Pareto Distribution (GPD), taking the speed limit as the threshold. However, this speed limit is an amount imposed by traffic authorities and may not necessarily represent an appropriate threshold according to Extreme Value Theory (EVT). According to Pickands (1975), excesses converge to the GPD as $u \to \infty$, and based on this, the GPD modeling is effective for a suitably high threshold value. Thus, the model proposed by Nascimento et al. (2012) allows the threshold to be a free parameter, with the data modeling itself indicating which portion is better represented by a mixture of distributions and which part is better modeled by the GPD. Nevertheless, for comparison purposes, modeling that considers all excesses derived from the GPD distribution was also conducted.

After estimating the parameters for the number of daily occurrences N and the parameters for speeding X considering for this the best model according to the Table 3, we can estimate the parameters of the Compound Poisson Process Y and the expected value through Equation (4). Specifically, in this work, as the value of the fine can increase for each speeding range, we will estimate the probability of each excess range associated with the amount to be paid within that range. This will allow us to obtain an estimate of the expected collection value.

3. Applications to Real Data

To illustrate the applicability of the proposed model, radar data were collected over a period of seven years, from 2015 to 2022, in the city of Teresina, PI, Brazil. The radars are located on four avenues in the city: Av Alameda Parnaíba (Alameda), Av Raul Lopes Shopping (Shopping), Av Maranhão (Maranhao) and Av Barão de Castelo Branco (Barao).

For each radar, an analysis of all excesses was initially performed using the $MGPD_k$ model from Nascimento et al. (2012). To choose the optimal number k of Gamma mixture components in the $MGPD_k$ model, the DIC (Spiegelhalter et al., 2002) criterion was used.

Table 3 presents the results for all models, including the use of a full GPD distribution and the $MGPD_k$ models with $k = 1$ and $k = 2$. For all radar speeding data, the $MGPD_2$ model was found to be the best fit.

Table 3: DIC fit measures.

Endereço	GPD	$k = 1$	$k = 2$
Alameda	263278.7	247517.2	218385.2
Shopping	267665.1	253704.1	210970.1
Maranhao	151864.8	145195.4	122570.4
Barao	168909.7	160787.6	157461.5

Table 4 presents the estimation of the model parameters. Note that in all cases, the estimation of the threshold u is found at a low value of excess speed, indicating that almost all excesses are modeled by the GPD. This was expected because we are analyzing speeding data, and the very nature of this data makes the GPD distribution suitable, with only a small portion being modeled by the Gamma mixture. Although the thresholds are low, considering this small portion of the data modeled by the Gamma mixture provided an advantage over treating all excesses as a GPD distribution, as indicated by the DIC in Table 3.

Considering the other parameters of the GPD distribution, we observe that the radars on Alameda and Barao avenues exhibit greater dispersion, with the parameter σ being larger than that of the radars on the other two avenues. Regarding the weight of the tail, the radar on Avenida Maranhao has observations with heavier tails, while Alameda and Shopping have lighter tails. For the radar on Barão Avenue, the estimate of $\xi \approx 0$ indicates that the speeding behavior at this location has an exponential tail.

Figure 2 illustrates the evolution of the estimated high quantiles for each radar. It is observed that on Shopping Avenue, the 95% and 99% quantiles exhibit the lowest levels of high excess quantiles, while the radar on Maranhao Avenue shows the highest values of excess in the high quantiles.

TABLE 4: Parameter estimates with 95% CI.

	Alameda			Maranhao			Shopping			Barao		
	Estimated	L.I.	L.S.	Estimated	L.I.	L.S.	Estimated	L.I.	L.S.	Estimated	L.I.	L.S.
ξ	-0.06	-0.07	-0.05	0.19	0.16	0.21	-0.04	-0.05	-0.04	0.00	-0.01	0.01
σ	6.33	6.18	6.45	4.25	4.17	4.35	5.40	5.32	5.46	6.20	6.09	6.31
u	3.01	3.01	3.01	2.00	1.99	2.00	3.00	3.00	3.00	3.01	3.01	3.02
μ_1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.02	1.00	1.09
μ_2	3.55	3.52	3.57	10.36	8.30	11.93	3.47	3.46	3.49	3.62	3.60	3.67
η_1	219.16	219.05	219.18	353.15	353.04	353.17	392.15	392.07	392.24	13.00	9.98	14.20
η_2	10.07	9.85	10.40	10.40	8.54	11.94	10.35	10.12	10.47	8.85	8.58	9.14
w_1	10.40	8.54	11.94	0.19	0.19	0.19	0.21	0.20	0.21	0.20	0.20	0.20
w_2	0.81	0.81	0.81	0.81	0.81	0.81	0.79	0.79	0.80	0.80	0.80	0.80

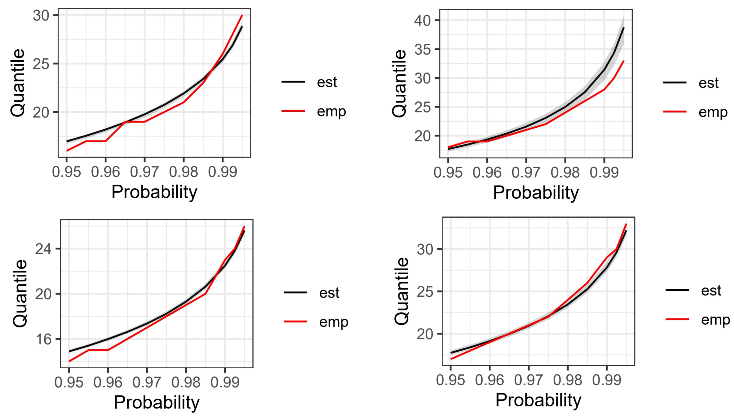

FIGURE 2: Quantile plot of the speedings: Estimated quantile (black line), with empirical quantile (red line) and 95% CI (grey area). Top: Alameda (Left) and Maranhao (Right). Bottom: Shopping (Left) and Barao (Right).

3.1. Estimating Revenue from Speeding Fines

Considering the Compound Poisson Process model, which represents the daily number of speedings using the Poisson distribution, and the value of each speeding violation through the $MGPD_k$ model, we present the estimation of the expected value of daily and annual tax collection from speeding fines. The concept of the expected value in the Compound Poisson Process will be used to estimate the expected value of daily collection.

Thus, considering Tables 1 and 2, the expected daily tax collection distribution for radars with speed limits of 40 km/h and 60 km/h can be calculated as follows:

$$A_{40} = \hat{\lambda}(130.16P(X \leq 8) + 195.23P(8 \leq X \leq 20) + 880.41P(X \geq 20))$$
$$A_{60} = \hat{\lambda}(130.16P(X \leq 12) + 195.23P(12 \leq X \leq 30) + 880.41P(X \geq 30)),$$

where $\hat{\lambda}$ is the estimated value of the number of speeding violations, calculated according to the posterior mean of λ in (5) for each radar. The probabilities of X is estimated according to the $MGPD_2$ distributions for each radar.

Table 5 presents the estimated value of daily collections for each radar. It concludes that the Shopping Avenue radar has the highest daily collection potential, with a value of 6383.88. Following this, Alameda Avenue has an estimated collection of 3727.48. Maranhao and Barao avenues have values that are closer to the 2000.00 range.

TABLE 5: Expected daily tax collection.

Endereço	IC -	IC +	Median
Alameda	3704.352	3764.832	3727.48
Shopping	6336.130	6427.172	6383.88
Maranhao	2657.423	2728.497	2693.00
Barao	2841.207	2900.338	2875.46

By multiplying the daily collection by the 365 days of the year, Table 6 presents the annual collection potential of each radar for fines. Notably, the Shopping Avenue radar is the only one to exceed the 2 million mark, with the potential of 2.32 million. Alameda and Barao surpass the 1 million barrier. Maranhao is restricted to an annual potential of just under 0.8 million.

TABLE 6: Expected annual tax collection.

Endereço	IC −	IC +	Median
Alameda	1352088.7	1374163.9	1360530
Shopping	2312687.5	2345918.0	2330116
Maranhao	969959.2	995901.6	982945
Barao	1037040.7	1058623.2	1049543

4. Concluding Remarks

This work was motivated by the need to predict revenue collection from specific radars, utilizing the Compound Poisson Process methodology within the context of extreme values.

By successfully modeling both the daily number of speed violations and the distribution of those violations, we were able to estimate the daily and annual revenue potential for the municipality. This information allows for better planning of expenses and revenues, ultimately enabling more effective organization of public traffic policies.

The model proposed in this work can also be applied in other fields, particularly in the insurance sector, where there is significant interest in analyzing both the value of claims and the number of claims occurring within a given period.

[Received: February 2024 — Accepted: October 2024]

References

Adelson, R. M. (1966), ‘Compound poisson distributions’, *Journal of the Operational Research Society* **17**, 73–75.

Ata, N. & Ozel, G. (2013), ‘Survival functions for the frailty models based on the discrete compound poisson process’, *Journal of Statistical Computation and Simulation* **11**, 2105–2116.

Casella, G. & Berger, R. L. (2001), *Statistical Inference*, Cengage Learning; 2nd ed.

Coles, S. G. (2001), *An Introduction to Statistical Modelling of Extreme Values*, Springer.

Embrechts, P., Küppelberg, C. & Mikosch, T. (1997), *Modelling Extremal Events for Insurance and Finance*, New York: Springer.

Fisher, R. A. & Tippett, L. H. C. (1928), 'On the estimation of the frequency distributions of the largest and smallest sumber of a sample', *Proceedings of the Cambridge Philosophical Society* **24**, 180–190.

Jenkinson, A. F. (1955), 'The frequency distribution of the annual maximum (or minimum) values of meteorological events', *Quarterly Journal of the Royal Meteorology Society* **81**, 158–171.

Nascimento, F. F., Gamerman, D. & Lopes, H. F. (2012), 'A semiparametric bayesian approach to extreme value estimation', *Statistics and Computing* **22**, 661–675.

Pickands, J. (1975), 'Statistical inference using extreme order statistics', *Annals of Statistics* **3**, 119–131.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & Linde, A. (2002), 'Bayesian measures of model complexity and fit', *Journal of the Royal Statistical Society B* **64**, 583–639.

von Mises, R. (1954), 'La distribution de la plus grande de n valeurs', *American Mathematical Society* **2**, 271–294.

Wiper, M., Rios Insua, D. & Ruggeri, F. (2001), 'Mixtures of gamma distributions with applications', *Journal of Computational and Graphical Statistics* **10**, 440–454.

Zhang, H., Liu, Y. & Li, B. (2014), 'Notes on discrete compound poisson model with applications to risk theory', *Insurance: Mathematics and Economics* **59**, 325–336.