Zero-inflated Truncated Generalized Pareto Distribution for the Analysis of Radio Audience Data

Couturier, Dominique-Laurent; Victoria-Feser, Maria-Pia

2009.01

FACULTE DES SCIENCES ECONOMIQUES ET SOCIALES

HAUTES ETUDES COMMERCIALES

ZERO-INFLATED TRUNCATED GENERALIZED PARETO DISTRIBUTION FOR THE ANALYSIS OF RADIO AUDIENCE DATA

D.-L. COUTURIER
M.-P. VICTORIA-FESER

HEC GENÈVE

UNIVERSITÉ DE GENÈVE

Zero-Inflated Truncated Generalized Pareto Distribution for the Analysis of Radio Audience Data

D.-L. Couturier and M.-P. Victoria-Feser

December 2008

University of Geneva

Abstract

Extreme value data with a high clump-at-zero occur in many domains. Moreover, it might happen that the observed data are either truncated below a given threshold and/or might not be reliable enough below that threshold because of the recording devices. This situations occurs in particular with radio audience data measured using personal meters that record environmental noise every minute, that is then matched to one of the several radio programs. There are therefore genuine zeroes for respondents not listening to the radio, but also zeroes corresponding to real listeners for whom the match between the recorded noise and the radio program could not be achieved. Since radio audiences are important for radio broadcasters in order for example to determine advertisement price policies, possibly according to the type of audience at different time points, it is essential to be able to explain not only the probability of listening a radio but also the average time spent listening the radio by means of the characteristics of the listeners. In this paper, we propose a generalized linear model for zero-inflated truncated Pareto distribution (ZITPo) that we use to fit audience radio data. Because it is based on the generalized Pareto distribution, the ZITPo model has nice properties such as model invariance to the choice of the threshold and from which a natural residual measure can be derived to assess the model fit to the data. From a general formulation of the most popular models for zero-inflated data, we derive our model by considering successively the truncated case, the generalized Pareto distribution and then the inclusion of covariates to explain the non-zero proportion of listeners and their mean listening time. By means of simulations, we study the performance of the maximum likelihood estimator (and derived inference) and use the model to fully analyze the audience data of a radio station in an area of Switzerland.

1 Introduction

Audience indicators – like rating[1], time spent listening[2] and market share – are essential for radio stations managers and advertisers. They give important indications on public profiles and on radio stations benchmarking allowing proper radio programming and optimization of advertising strategies. The weaknesses of traditional audience measurements methods based on individual recollection of the time spent listening to all radio stations led to the development of individual, portable, and passive electronic measurement systems providing more reliable and detailed measures (Heindervckx and Phillips 2001). Radiocontrol[3] developed a “wristwatch meter”, which records 4 seconds of ambient sound every minutes and compares this sequence to the corresponding one of all available radios. The “people portable meter” of Arbitron[4] or the “Eurisko multimedia monitor” of Gfk[5] consist in a pager-sized device which detects inaudible codes that broadcasters embed in their programs.

Hence, the fundamental audience measure available through these portable and passive measurement systems is a dichotomous variable Y_{ismt} indicating if the participant i was listening to the radio station s at the minute m of the day t. Most used audience indicators for a given radio station are all functions of the sum of those quantities over a day part, mostly 24 hours, i.e. $Y_{ist} = \sum_{m=1}^{1440} Y_{ismt}$.

We have at our disposal radio audience data of the swiss measurement system ‘Radiocontrol’ in 2007. As illustrated in Figure 1, the distribution of the daily number of listening minutes for a given radio is extremely skewed, left-truncated and zero-inflated. In other words, firstly, the empirical distribution of the data

[1] Percentage of people who tune to a given radio station during a day.

[2] Average listening time to a given radio station per listener.

[3] http://www.radiocontrol.ch

[4] http://www.arbitron.com

[5] http://www.gfk.com

appears monotonically decreasing. The probability to listen to a radio during a time interval decreases with the time interval length. Secondly, because of contact validation rules of the swiss measurement system, listening times inferior to 3 minutes are recorded as zeroes. It should also be noted that a part of these observed zeroes as well as a part of the observed 3 to 4 minutes listening times may respectively be false zeroes or false positive observations. The probability to observe false positive or negative contact is negligible over a time interval of five or more consecutive minutes. Thirdly, the data contains a high clump-at-zero corresponding to the percentage of people that had no contact (or a contact of less than 3 to 5 minutes) with that radio station.

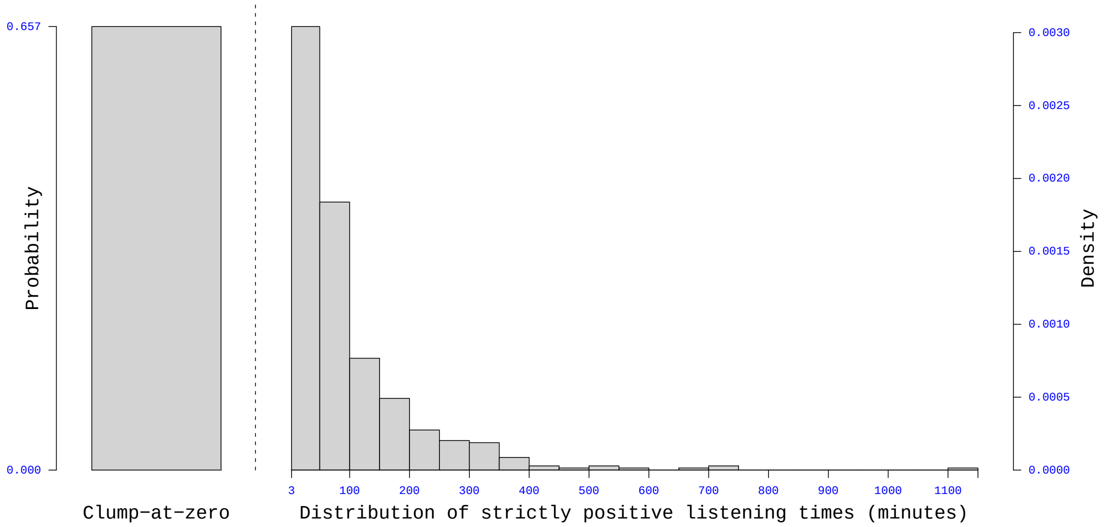

Figure 1: *Empirical distribution of the daily listening times to a national radio in an area of the French part of Switzerland during the first semester 2006. 1382 participants were measured by means of the Radiocontrol system during one day of the period of interest. Zeroes represent 65.7% of the data. The distribution of the positive data is extremely skewed with a maximum daily listening time of 1136 minutes. The lowest possible positive listening time is of 3 minutes.*

Data with a clump-at-zero and an asymmetric heavy-tail distribution occur in numerous disciplines. Examples are the daily levels of precipitation in an area (Weglarczyk et al. 2005), the yearly amount of car insurance claims per client

(Chapados et al. 2002; Christmann 2004) or the length of overnight stays at hospital per patient (Chen et al. 2007). However, no model has been proposed sofar for data with a clump-at-zero together with a truncation of small values under a threshold, a model that is necessary to describe, in particular, radio audience data like in our example, but also any other type of data that might, for example for recording reasons, have unreliable measurements at small values of the variable of interest. Hence, the purpose of this paper is to develop a model able to fit zero-inflated truncated heavy-tail data and to explain, by means of covariates, both the probability associated with a non-zero value and the expectation of positive outcomes. Such a model makes particularly sense in the context of radio audience: The probability of a non-null value and the expectation of positive outcomes respectively correspond to the rating and time spent listening audience indicators. Market shares are a function of these expectations.

Models for zero-inflated data have received a quite large attention in the literature. The most popular ones include the two-parts model of Duan et al. (1983) and the zero-inflated count models initiated by Lambert (1992) for continuous data, or the hurdle model of Mullahy (1986) for count data. In section 2, we describe our model as a natural extension of these models that takes into account the left truncation of the outcome variable. To model the positive part of the radio listening times, we propose a zeromodal Pareto-like distribution. Choice has been made for the generalized Pareto distribution because of its ability to fit heavy tails, to be "model invariant" to the choice of the threshold for the left truncation, and because it can be used to only model the tail of the distribution. The resulting model we propose is hence a zero-inflated truncated Pareto (ZITPo) model in which the probability of non-zero outcomes and the mean of the positive outcomes is linked to a set of covariates in a generalized linear model framework. The ZITPo has a great fitting flexibility and useful properties as argued in section

2.4. In section 3 we investigate by means of simulations the sample properties of the maximum likelihood estimator and inferential procedures. Since ZITPo models are new, it is also important to be able to check the fit of the model and therefore we propose in section 4 a new data analysis tool based on Pareto residuals that is derived in a natural manner from the properties of the ZITPo model. The data from a radio station in an area of Switzerland are then fully analyzed in section 5 by means of the ZITPo which provides and excellent fit to the data and hence good explanatory power for the probability of non-zero outcomes and the mean of the positive outcomes.

2 The ZITPo model

The generalized Pareto distribution, introduced by Pickands (1975), is a limit distribution for the excess over a (large) threshold α for data coming from generalized extreme value distributions, as well as a generalization of the Pareto distribution. The three-parameters generalized Pareto distribution has the following cumulative distribution function:

$$F_Y(y|\alpha,\tau,\xi) = \begin{cases} 1-(1+\xi\frac{y-\alpha}{\tau})^{-\frac{1}{\xi}} & \text{if } \xi \neq 0, \\ 1-\exp(-\frac{y-\alpha}{\tau}) & \text{if } \xi = 0; \end{cases} \tag{1}$$

where α, τ and ξ are location, scale and shape parameters, $\alpha \geq 0$ and $\tau > 0$. The range of y is $]\alpha, -\frac{\tau}{\xi+\alpha}[$ if $\xi < 0$, and $]\alpha, \infty[$ otherwise. Special cases are the exponential distribution with mean τ for $\xi = 0$, and the uniform distribution for $\xi = -1$. Pareto-like distributions occur for $\xi > 0$. The generalized Pareto distribution has been widely used to model rare events in several fields. Applications for environmental extremes are especially numerous (river flow, ozone levels, earthquakes).

For modeling audience radio data, it is also important to be able to link moments or parameters of the generalized Pareto distribution to a set of explanatory variables. The generalized linear models (GLM) framework, introduced by Nelder and Wedderburn (1972), provides a general setting to achieve this aim. GLM are a generalization of the linear regression model in which the assumption of normality of the conditional distribution of the responses vector $\mathbf{y}$ given a set of covariates $\mathbf{X}$, $\mathbf{y}|\mathbf{X}$, is relaxed. These models assume that the ith unit response, y_i, follows a distribution belonging to the exponential family, and the expectation of the ith response, y_i, is linked to a set of fixed covariates $\mathbf{x}_i$ through an invertible linear predictor function $\nu(\cdot)$, by means of $\mathrm{E}[Y_i] = \nu^{-1}(\mathbf{x}_i\boldsymbol{\beta})$, with $\boldsymbol{\beta}$

a set of regression coefficients. The generalized Pareto distribution falls outside the exponential family framework, and hence the advantages associated with this framework – like well known iterative estimation procedures and mathematical properties – are not available. However, extension of the GLM to distributions outside the exponential family is pretty straightforward.

Actually, generalized linear modeling exists since a long time with responses following extreme value distributions, but not in the traditional scheme that directly relates the response expectation to the explanatory variables through a linear predictor. Indeed, in extreme value analyzes, very often the parameters of the response distribution instead of the response expectation are linked to the covariates. Davison and Smith (1990, p. 395) consider that this represents "a more fruitful approach" than the usual one that links the distribution moments to the regressors, as the moments of generalized extreme value distribution do not exist for all values of their parameters. Following this approach, Chavez-Demoulin and Davison (2005) adapt generalized additive models to the generalized Pareto distribution to fit meteorological and environmental extremes. We refer to Coles (2001, section 6.4) for a review. In survival analysis, depending of the choice of the hazard function $h(t)$, the survival function $f(t)$ may follow an extreme value distribution. In this context, the hazard function $h(t) = \frac{f(t)}{1-F(t)}$, is then related to the covariates through a linear predictor instead of the response expectation. Such developments may be found in Aitkin and Clayton (1980). As we will see in more details below, for the purpose of modeling radio audience data, it is more sensible to link the excepted value of the response to a set of covariates.

Before adapting the generalized Pareto distribution to handle clump-at-zero and left truncation of the positive part of the data, as well as incorporating in the resulting model covariates in order to explain the probability of a zero outcome and the mean of the positive part, we briefly describe models proposed sofar for zero-inflated data. The aim is to propose a general formulation from which different models for different situations can be deduced, and in particular from which we build our zero inflated truncated Pareto (ZITPo) model. We then also describe in details the ZITPo model assumptions and discuss some possible extensions.

2.1 Models for nonnegative zero-inflated data

There is a rich literature about adaptation of statistical models to the case of zero-inflated data. We refer to Min and Agresti (2002, 2005) and Ridout et al. (1998) for a review. Min and Agresti (2002) compare the advantages and disadvantages of existing approaches and note that the most appealing modeling for zero-inflated continuous data is the two-parts model of Duan et al. (1983), and the zero-inflated count models initiated by Lambert (1992) or the hurdle model of Mullahy (1986) in the case of count data with a clump-at-zero.

These models are similar. Their key idea is to mix two random variables: A first one, Y_1, that handles the zeroes excess and a second one, Y_2, that models the other part of the data. Y_1 typically follows a Bernoulli distribution where $P_{Y_1}(0) = 1 - \pi$ denotes the probability to observe a zero outcome. In the hurdle and two-parts models (also called conditional models), the probability of the data being equal to zero only depends on Y_1 and the positive data are all modeled by Y_2, which may follow a zero-truncated distribution in the case of count data (hurdle model) or a continuous distribution (two-parts model). In these cases, $P_{Y_2}(0) = 0$. In zero-inflated models (also called mixture models), Y_2 does not follow a zero-truncated distribution. The probability associated to zero thus depends on both Y_1 and Y_2.

Let Y be a random variable with probability distribution P_Y for the clump-at-zero and the positive part, when the latter is discrete, i.e. Y_2 is discrete, then P_Y may be expressed in the following way:

$$P_Y(y) = \Big[P_{Y_1}(0)+(1-P_{Y_1}(0))P_{Y_2}(y)\Big]\iota(y=0)+\Big[(1-P_{Y_1}(0))P_{Y_2}(y)\Big]\Delta_0(y), \quad (2)$$

where $y = 0, 1, 2, ...$, $\iota(\cdot)$ is the indicator function which equals one if the condition is true and zero otherwise, and $\Delta_0(y)$ is a step function taking the value of one for $y > 0$ and zero otherwise. When Y_2 is continuous or semicontinuous (see Min and Agresti 2002, p. 7), we have the following density function for Y:

$$f_Y(y) = \Big[P_{Y_1}(0) + (1 - P_{Y_1}(0))P_{Y_2}(0)\Big]\delta(y) + \Big[(1 - P_{Y_1}(0))f_{Y_2}(y)\Big]\Delta_0(y) \quad (3)$$

where $\delta(y)$ is a Dirac delta function which equals zero for $y \neq 0$, and $y \in [0, \infty[$. Note that when $P_{Y_2}(0) = 0$, we have the hurdle or two parts models, while we have zero-inflated models when this is not the case.

The use of the generalized Pareto distribution to model zero-inflated data is not common, one exception being Weglarczyk et al. (2005). The authors compare the fitting ability of some semicontinuous distributions to fit zero-inflated hydrological data and consider a Dirac generalized Pareto distribution with density function

$$f_Y(y|\pi, \tau, \xi) = (1 - \pi)\delta(y) + \frac{\pi}{\tau}\left(1 + \xi\frac{y}{\tau}\right)^{-\frac{1}{\xi}-1}\Delta_0(y), \quad (4)$$

where $\tau > 0$, $\xi \neq 0$, $0 \leq (1-\pi) \leq 1$ corresponds to the probability of a zero event. Note that compared to (1), $\alpha = 0$. The Dirac generalized Pareto distribution in (4) thus corresponds to a two-parts model with $P_{Y_2}(0) = 0$ in which $f_{Y_2}(y)$ is the density function of the generalized Pareto distribution.

In the following sections, we propose to extend (3) (and (4)) to take into account the possible truncation of small values as well as to incorporate covariates to explain (a function of) the probability of zero outcomes and the mean distribution of positive outcomes.

2.2 The ZITPo distribution

Let Y^* denote the effective (but unknown) daily listening time for a given radio. The probability and cumulative distribution functions of Y^*, $f_{Y^*}(y^*)$ and $F_{Y^*}(y^*)$, are semicontinuous with a point mass in zero and a continuous distribution for the positive values. Let Y denote the observed listening times with density function $f_Y(y)$. As listening times inferior to a given value y° are recorded as zeroes, observed zeroes are then a mixture between the effective zero listening times and the positive listening times reported as zeroes because of the measurement system. Accordingly $F_Y(0) = F_{Y^*}(y^\circ)$.

A semicontinuous version of the zero-inflated count model described in (3) is indeed adequate to model the double origins of the zeroes in the clump-at-zero and the positive values of the observed listening times. Let assume that the unknown and true proportion of zero listening times is $1 - \pi$, with $0 \leq \pi \leq 1$, and that the effective positive listening times follow a two-parameters generalized Pareto distribution (with $\alpha = 0$), $Y^*|(Y^* > 0) \sim \text{GPD}(\tau, \xi)$. Then, in (3), $P_{Y_1}(0) = 1 - \pi$ corresponds to the effective proportion of non-listeners, and $P_{Y_2}(0) = F_{(Y^*|Y^*>0)}(y^\circ)$ corresponds to the part of the two-parameters generalized Pareto distribution that can't be observed because of the measurement system limitations. The density functions of the effective listening times Y^* and of the observed listening times Y are

$$\begin{aligned} f_{Y^*}(y^*|\pi,\tau,\xi) &= \Big[1-\pi\Big]\delta(y^*) + \left[\frac{\pi}{\tau}\left(1+\xi\frac{y^*}{\tau}\right)^{-\frac{1}{\xi}-1}\right]\Delta_0(y^*), \qquad (5) \\ f_Y(y|\pi,\tau,\xi) &= \Big[(1-\pi)+\pi F_{(Y^*|Y^*>0)}\left(y^\circ\right)\Big]\delta(y) + \Big[\pi f_{(Y^*|Y^*>0)}(y)\Big]\Delta_{y^\circ}(y) \\ &= \left[1-\pi\left(1+\xi\frac{y^\circ}{\tau}\right)^{-\frac{1}{\xi}}\right]\delta(y) + \left[\frac{\pi}{\tau}\left(1+\xi\frac{y}{\tau}\right)^{-\frac{1}{\xi}-1}\right]\Delta_{y^\circ}(y), \qquad (6) \end{aligned}$$

where $0 \leq \pi \leq 1$, $\tau > 0$, $\xi \neq 0$ and $y^\circ \geq 0$. For $y^\circ = 0$, (6) reduces to the Dirac generalized Pareto described in (4). Finally note that if the observed listening times distribution in (6) has the disadvantage of being a mixture distribution which makes it more complex to fit, its underlying distribution in (5) takes the advantages of the orthogonal parameterization of the hurdle and two stages models and is thus easier to interpret (for a discussion on the orthogonal parameterization see e.g. Welsh et al. 1996). Indeed the zeroes depend on π, while the positive outcomes rely on the generalized Pareto parameters, τ and ξ.

Figure 2 shows the distribution of a dataset simulated from a ZITPo distribution. The theoretical untruncated and truncated distribution functions, respectively corresponding to (5) and (6), are superimposed to the plot in black and red lines. On the discrete part of the plot, the surfaces within the red and black boxes correspond to the theoretical probabilities to observe zeroes when there is (red) and when there is no (black) left truncation of the positive part of the data. Those probabilities respectively equal $1-\pi$ and $(1-\pi)+\pi F^{-1}_{(Y^*|Y^*>0)}(y^\circ)$. On the continuous part of the plot, the expectations of the truncated (red) and untruncated (black) distributions are indicated. It is then clear that the expected value for the true listening time Y^* (given by μ in black in Figure 2) is different from the expected value of the truncated distribution (given by μ in red in Figure 2). For the audience data, one quantity of interest is μ for the untruncated distribution.

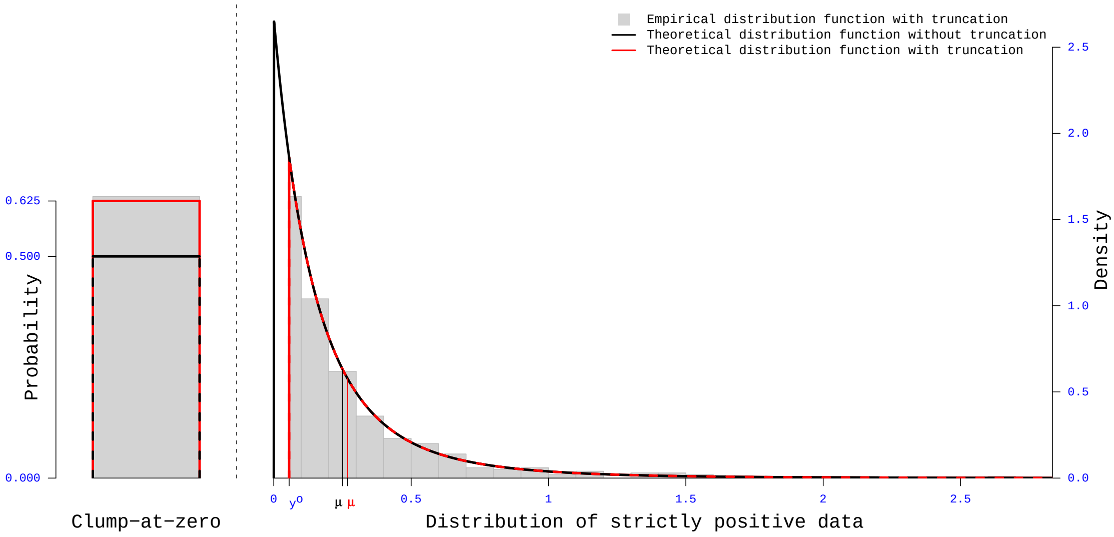

Figure 2: *Empirical distribution function of a dataset simulated from a ZITPo model with parameters* $\pi = 0.5$*,* $\mu = \xi = 0.25$ *and* $y^\circ = F^{-1}_{(Y^*|Y^*>0)}(0.25)$*. The theoretical truncated and untruncated densities functions are superimposed to the plot with red and black lines. The value of the expectations of the positive values of the truncated and untruncated distributions are indicated on the x-axis. On the discrete part of the plot, the surfaces within the red and black boxes correspond to the theoretical probabilities to observe zeroes when there is (red) and when there is no (black) left truncation of the positive part of the data. Those probabilities respectively equal* $1 - \pi$ *and* $(1 - \pi) + \pi F^{-1}_{(Y^*|Y^*>0)}(y^\circ)$

2.3 Covariates modeling in ZITPo distribution

Adaptation of the GLM to zero-inflated models is very intuitive. The expectations of the distributions of Y_1 and Y_2 in (2) and (3) are linked to the covariates through adapted link functions. The logit link is often chosen to relate the expectation of Y_1, corresponding to the probability to observe positive values, to the covariates. The log link makes sense to connect the expectation of Y_2, corresponding to the mean of the positive data, to the covariates, as this last is necessarily positive. For the ith observation, we then have

$$\pi_i = \mathrm{P}(Y_i > 0) = \nu_1^{-1}(\mathbf{x}_{i1}^T\boldsymbol{\beta}_1) = \frac{\exp(\mathbf{x}_{i1}^T\boldsymbol{\beta}_1)}{1 + \exp(\mathbf{x}_{i1}^T\boldsymbol{\beta}_1)}, \tag{7}$$

$$\mu_i = \mathrm{E}[Y_i|Y_i > 0] = \nu_2^{-1}(\mathbf{x}_{i2}^T\boldsymbol{\beta}_2) = \exp(\mathbf{x}_{i2}^T\boldsymbol{\beta}_2), \tag{8}$$

where $\nu_1^{-1}(\cdot)$ and $\nu_2^{-1}(\cdot)$ are the inverse of the linear predictor functions linking the expectations of Y_1 and Y_2 to the covariates, $\mathbf{x}_{i1}$ and $\mathbf{x}_{i2}$ are the covariates of the ith observation that may contain the same predictors, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are the corresponding parameters.

Inclusion of covariates in (4) requires to express the distribution $f_Y(y)$ in terms of the expectation of the positive values of the data. Let $(Y^*|Y^* > 0) \sim \text{GPD}(\tau, \xi)$. Then

$$\mu = \mathrm{E}\big[Y^*|Y^* > 0\big] = \frac{\tau}{1-\xi} \text{ for } 1-\xi > 0.$$

The first moment of the generalized Pareto distribution, μ, thus exists for values of ξ lower than one. Substituting τ by $\mu(1-\xi)$ in (6) gives

$$\begin{aligned} f_Y(y|\pi,\mu,\xi) = {} & \left[1 - \pi\left(1 + \left(\frac{\xi}{1-\xi}\right)\frac{y^\circ}{\mu}\right)^{-\frac{1}{\xi}}\right]\delta(y) \\ & + \left[\frac{\pi}{\mu(1-\xi)}\left(1 + \left(\frac{\xi}{1-\xi}\right)\frac{y}{\mu}\right)^{-\frac{1}{\xi}-1}\right]\Delta_{y^\circ}(y), \end{aligned} \tag{9}$$

with $0 \le \pi \le 1$, $\mu > 0$, $\xi \ne 0$ and $\xi < 1$, $y^\circ \ge 0$. The inclusion of the covariates as described in (7) and (8) is now straightforward. For the ith observation, we have

$$\begin{aligned} & f_{Y_i}(y_i|\mathbf{x}_{i1}, \mathbf{x}_{i2}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \xi) = \\ & \left[1 - \frac{\exp(\mathbf{x}_{i1}^T\boldsymbol{\beta}_1)}{1+\exp(\mathbf{x}_{i1}^T\boldsymbol{\beta}_1)}\left(1 + \left(\frac{\xi}{1-\xi}\right)\frac{y^\circ}{\exp(\mathbf{x}_{i2}^T\boldsymbol{\beta}_2)}\right)^{-\frac{1}{\xi}}\right]\delta(y) + \\ & \left[\frac{\exp(\mathbf{x}_{i1}^T\boldsymbol{\beta}_1)}{1+\exp(\mathbf{x}_{i1}^T\boldsymbol{\beta}_1)}\frac{1}{\exp(\mathbf{x}_{i2}^T\boldsymbol{\beta}_2)(1-\xi)}\left(1 + \left(\frac{\xi}{1-\xi}\right)\frac{y_i}{\exp(\mathbf{x}_{i2}^T\boldsymbol{\beta}_2)}\right)^{-\frac{1}{\xi}-1}\right]\Delta_{y^\circ}(y). \end{aligned} \tag{10}$$

2.4 Assumptions, properties and extensions of ZITPo models

The form of the ZITPo model implies a number of assumptions on the distribution of the positive values. For example, the unobserved positive listening times belonging to the range $]0, y^\circ[$ correspond to the non-observed part of a left-truncated generalized Pareto distribution. As the generalized Pareto density function is zeromodal and monotonically decreasing, this assumption implies that, conditionally on the covariates, the probability of positive listening times in the interval $]0, y^\circ[$ is higher than in any other interval of the same size. As zapping through radio is frequent, we believe that this assumption is realistic.

Moreover, conditionally on the covariates, the real positive listening times follow generalized Pareto distributions having different expectations μ_i but sharing the same ξ-value: $Y_i^*|(Y_i^* > 0) \sim \text{GPD}(\mu_i, \xi)$. We can observe that

$$F_{(Y_i^*|Y_i^*>0)}(\mu_i) = 1 - \left(1 + \xi \frac{\mu_i}{\mu_i(1-\xi)}\right)^{-\frac{1}{\xi}} = 1 - (1-\xi)^{\frac{1}{\xi}}, \tag{11}$$

with $\xi \neq 0$. An assumption of this model is thus that the expectation μ_i always corresponds to the quantile $1-(1-\xi)^{\frac{1}{\xi}}$ of a $\text{GPD}(\mu_i, \xi)$. Figure 3 shows examples of two-parameters generalized Pareto density functions sharing the same ξ-value (within the same graph) but having different expectations. For the same ξ-value, the density functions show a great variety of forms and thus a high ability to model different datasets with more or less heavy tails.

One should also stress that because of the reparametrization of the generalized Pareto density formulated in (9), the shape parameter is restricted to values lower than one. This doesn't seem problematic in regard to (11). Indeed, for $\xi > 0.95$, μ corresponds to quantiles of the distribution higher than 0.95. We don't expect cases in which the theoretical mean belongs to the last 5% of the distribution at least with radio listening data.

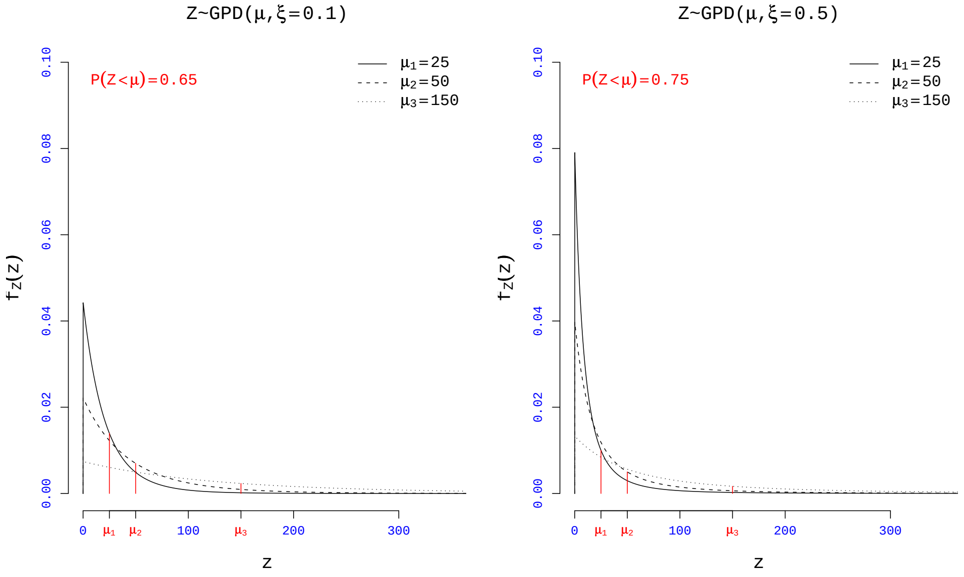

Figure 3: *Examples of two-parameters generalized Pareto distributions. In both plots, three distribution functions sharing the same ξ-value are proposed. Their respective expectations are $\mu_1 = 25, \mu_2 = 50$ and $\mu_3 = 100$. The probability to observe data below the expectation is indicated in red.*

Even if there are some restrictions in the use of ZITPo models, the two-parts form of the density described in (10) as well as properties of the generalized Pareto distribution offer to ZITPo models additional abilities to fit and analyze a variety of datasets, in particular our radio audience data in Switzerland. For example, in addition to the problem of truncation of the listening times lower than 3 minutes, the observed listening times are not completely reliable in the neighborhood of the truncation boundary. Indeed, some of the 3 to 4 minutes effective positive listening times are coded as zeroes and some effective zeroes correspond to small observed positive listening times. An interesting property of ZITPo models is that y° may be chosen such that the observed data lower than y° integrate the most part of the false zero and false positive observations. If all observed positive data inferior to y° are coded as zeroes in order to belong to the clump-at-zero in (6), the model will estimate the parameters of $f_{Y^*}(y^*|\pi, \tau, \xi)$ without being affected by the errors of the measurement system occurring on $[0, y^{\circ}[$.

Moreover, for radio stations managers and advertisers, the behavior of important radio listeners is of particular interest as they represent the core of their audience. An appreciated information is the average listening time of people listening to a radio station more than $y^{\bullet}$ minutes (typically a value over the median of the positive outcomes), that is $\mathrm{E}[Y^{*}|Y^{*} > y^{\bullet}]$. The stability with respect to excess over threshold operations of the generalized Pareto distribution (see e.g. Castillo and Hadi 1997, p. 1610, or Coles 2001, p. 79) and the shifting property of distribution of the location family, allow to easily determine the distribution of the data over a threshold $y^{\bullet}$. Let $Y_i^{*+} = (Y_i^{*}|Y_i^{*} > 0)$ denote the positive values of the model, with $Y_i^{*+} \sim \mathrm{GPD}(\tau_i, \xi)$ with $\tau_i = \mu_i(1-\xi)$. Then we have that

$$f_{(Y_i^{*+}|Y_i^{*+}>y^{\bullet})}(y_i^{*+}|\tau_i,\xi) = \frac{1}{\tau_i - \xi y^{\bullet}}\left(1 + \xi\frac{y_i^{*+} - y^{\bullet}}{\tau_i - \xi y^{\bullet}}\right)^{-\frac{1}{\xi}-1}. \qquad (12)$$

The distribution of the listening time over a threshold thus follows a three-parameters generalized Pareto distribution of parameters $\alpha^{\bullet} = y^{\bullet}$, $\tau_i^{\bullet} = \tau_i - \xi y^{\bullet}$ and $\xi^{\bullet} = \xi$. The corresponding expected listening time over a threshold of $y^{\bullet}$ minutes, $\mu_i^{\bullet}$, is then given by

$$\mu_i^{\bullet} = \mathrm{E}[Y^{*+}|Y^{*+} > y^{\bullet}] = \frac{\tau_i^{\bullet}}{1-\xi} + y^{\bullet} = \mu_i + \frac{\xi y^{\bullet}}{1-\xi} + y^{\bullet}, \qquad (13)$$

where $\mu_i = \mu$ in simple models without covariates and $\mu_i = \exp(\mathbf{x}_{i2}^T\boldsymbol{\beta}_2)$ in models incorporating covariates. The expectation of the positive data over a threshold (i.e. the expectation of the data on $]y^{\bullet}, \infty[$) thus simply corresponds to a linear shift of the expectation of the positive data on $]0, \infty[$. There is therefore no need to change the ZITPo model when one is interested in $\mu_i^{\bullet}$, or in oder words, the effect of the covariates on $\mu_i^{\bullet}$ is the same as on μ_i.

Finally, in the same spirit as above, the ZITPo model can easily be extended to the three-parameters generalized Pareto distribution by introducing a shift parameter $y^\bullet \leq y^\circ$ corresponding to α in (1). In (9) and (10) we have that $y^\bullet = 0$. Adding the shift parameter makes sense if information below $y^\bullet$ is not of direct interest, like if no listeners and listeners that only zap through a given radio are considered alike for the radio broadcaster. The resulting model which extends (9) (and consequently (10)) would allow to model the probability to get an outcome lower than a given positive value $y^\bullet$ as well as the expectation of the data over $y^\bullet$, with positive outcomes observed above y°. In this case, all data lower than $y^\bullet$ would be treated as "zeroes" in order to be part of the clump-at-zero. The density functions of the observed listening times Y would then be (for an observation y_i)

$$\begin{aligned} f_Y(y_i|\pi_i^\bullet, \mu_i^\bullet, \xi^\bullet) = & \left[1 - \pi_i^\bullet \left(1 + \left(\frac{\xi}{1-\xi}\right)\left(\frac{y^\circ - y^\bullet}{\mu_i^\bullet - y^\bullet}\right)\right)^{-\frac{1}{\xi}}\right]\delta(y_i) \\ & + \left[\frac{\pi_i^\bullet}{(\mu_i^\bullet - y^\bullet)(1-\xi)}\left(1 + \left(\frac{\xi}{1-\xi}\right)\left(\frac{y_i - y^\bullet}{\mu_i^\bullet - y^\bullet}\right)\right)^{-\frac{1}{\xi}-1}\right]\Delta_{y^\circ}(y_i). \end{aligned} \tag{14}$$

The parameters $\pi_i^\bullet$ and $\mu_i^\bullet$ can possibly be linked to a set of covariates as in done in (10). If there is no y°-truncation and if the data on $]y^\bullet, y^\circ[$ are reliable, $y^\circ = y^\bullet$ and (14) is reduced to a two-stages model since the first part of the right handside of (14) reduces to $(1 - \pi_i^\bullet)\delta(y_i)$. This extension is particularly useful when the interest only lies on the tail distribution of the positive outcomes. Indeed, in that case $\pi^\bullet$ is a nuisance parameter and the generalized Pareto distributional assumption on $]0, y^\bullet[$ is no more necessary. For the model to fit the data (observed above y°), one only needs the assumption that the generalized Pareto distribution holds above $y^\bullet$, with a mean that possibly depends on a set of covariates and constant ξ. This might be an interesting setting for example in finance when seeking to explain the value-at-risk of financial instruments. In these cases however, the choice of $y^\bullet$ might become an important issue and criteria based on mean squared errors (see e.g. Hill 1975; Hall and Welsh 1985; Beirlant et al. 1996) or prediction errors (Dupuis and Victoria-Feser 2006) could in principle be extended to the ZITPo. In what follows, we will however focus on models with $y^{\bullet} = 0$.

3 Estimation and inference

Fitting methods for the generalized Pareto distribution in (1) (i.e. without a clump-at-zero) has been of great interest in the literature. Castillo and Hadi (1997) and Singh and Ahmad (2004) propose a comparative evaluation of the most used classical estimators for the two and three-parameters distributions. Robust estimators have also been developed (Dupuis and Tsao 1998; Peng and Welsh 2001; Juárez and Schucany 2004). We propose here to use the maximum likelihood estimator (MLE).

The parameters log-likelihood given the data and known information of the ZITPo model described in (10) is

$$
\begin{aligned}
&l(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \xi | \mathbf{y}, y^{\circ}, \mathbf{X}_1, \mathbf{X}_2) = \\
&\left\{ \sum_{i=1}^{n} \iota(y_i = 0) \log \left[1 - \frac{\exp(\mathbf{x}_{i1}^T \boldsymbol{\beta}_1)}{1 + \exp(\mathbf{x}_{i1}^T \boldsymbol{\beta}_1)} \left(1 + \left(\frac{\xi}{1 - \xi} \right) \left(\frac{y^{\circ}}{\exp(\mathbf{x}_{i2}^T \boldsymbol{\beta}_2)} \right) \right)^{-\frac{1}{\xi}} \right] \right\} + \\
&\left\{ \sum_{i=1}^{n} \Delta_{y^{\circ}}(y) \left[\mathbf{x}_{i1}^T \boldsymbol{\beta}_1 - \mathbf{x}_{i2}^T \boldsymbol{\beta}_2 - \log \left(\frac{(1 - \xi)^{-1}}{1 + \exp(1 + \mathbf{x}_{i1}^T \boldsymbol{\beta}_1)} \right) \right] \right\} + \\
&\left\{ \sum_{i=1}^{n} \Delta_{y^{\circ}}(y) \left(-\frac{1}{\xi} - 1 \right) \log \left(1 + \left(\frac{\xi}{1 - \xi} \right) \left(\frac{y_i}{\exp(\mathbf{x}_{i2}^T \boldsymbol{\beta}_2)} \right) \right) \right\}
\end{aligned}
\tag{15}
$$

Maximization of this expression is achieved using the quasi-Newton method with numerically computed gradient matrix. Convergence is obtained rapidly for most of the cases we have tried, even with models embedding many covariates. The use of slightly different starting values did always provide a solution to the unusual cases in which we met convergence problems. The program is implemented in R functions available upon request from the authors.

In order to check the finite sample properties of the MLE for the ZITPo model, we perform a simulation study. The first set of simulations focuses on simple models in which all the observations follow a ZITPo distribution like described in (9). In the second set of simulations we consider more complex models incorporating covariates as in (10).

For each simulation set, the MLE is computed on samples with three different sample sizes of respectively 500, 1000 and 2000 observations, simulated with two different values for the shape parameter, $\xi = 0.25$ and $\xi = 0.5$. The sampling distribution of the MLE are presented by means of boxplots on 2500 simulated datasets. Horizontal red lines indicate the position of the true parameter values. Blue percentages indicate the coverage levels of 95%- confidence intervals of the form $[\hat{\theta} - \Phi^{-1}(0.975)\hat{\sigma}_{\hat{\theta}}, \hat{\theta} + \Phi^{-1}(0.975)\hat{\sigma}_{\hat{\theta}}]$, where Φ is the density function of the standard normal distribution and where $\hat{\sigma}_{\hat{\theta}}$ are obtained from the inverse of the estimated hessian matrix.

For the first set of simulations, the observations were simulated from a ZITPo distribution with parameters $\pi = 0.5$, $\mu = 0.25$. The first quartile of a GPD($\mu = 0.25, \xi$), $F^{-1}_{(Y^*|Y^*>0)}(0.25)$, was chosen as the cutting value y°. The clump-at-zero represents 5/8 of the distribution. Figure 2 presents the distribution of one simulated dataset of 1000 observations with $\xi = 0.25$ and Figure 4 presents the results for simulations with $\xi = 0.25$ (similar results were obtained for $\xi = 0.5$).

Regardless of the sample size, the boxplots of the MLE of π and μ are well centered around the true parameters value. The coverage levels of the corresponding confidence intervals are close to the 95% nominal value. The left truncation

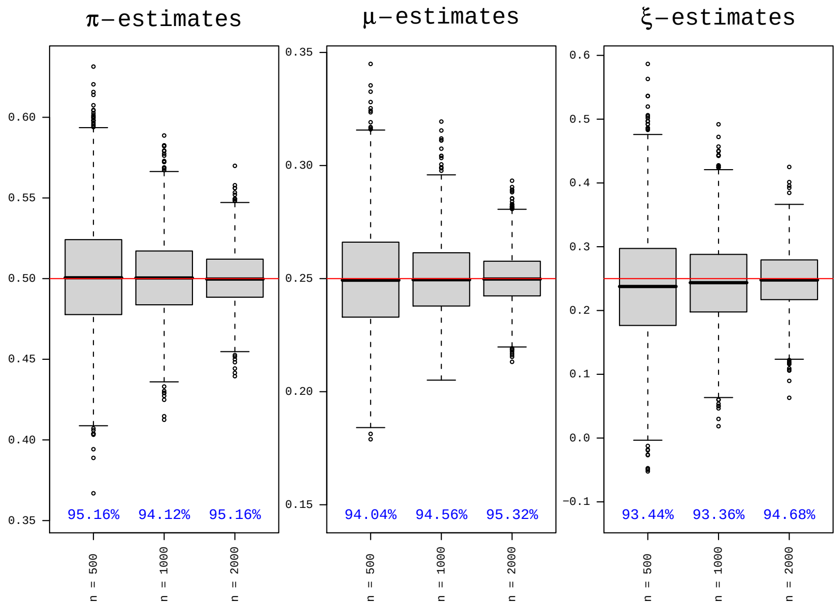

Figure 4: *Boxplots of the MLE computed on 2500 datasets simulated from a ZITPo distribution with parameters* $\pi = 0.5$*,* $\mu = 0.25$*,* $\xi = 0.25$ *and* $y^{\circ} = F^{-1}_{(Y^*|Y^*>0)}(0.25)$*. Analyzes were performed for samples of 500, 1000 and 2000 observations. The horizontal red lines indicate the position of the true parameter values. The blue percentages indicate the coverage levels of confidence intervals of the form* $[\hat{\theta} - \Phi^{-1}(0.975)\hat{\sigma}_{\hat{\theta}}, \hat{\theta} + \Phi^{-1}(0.975)\hat{\sigma}_{\hat{\theta}}]$*, where* Φ *is the density function of the standard normal distribution.*

of 25% of the generalized Pareto distribution thus doesn't prevent the MLE to adequately estimate the true proportion of zeroes and the true mean of the untruncated positive values.

Estimation of the shape parameter is known to be problematic even with large sample sizes and regardless of the estimating method (Hosking and Wallis 1987). The boxplots of the MLE of ξ in Figure 4 show a slight bias with $n = 500$, but no bias for greater values of n. This also confirms the findings of Chavez-Demoulin and Davison (2005, p. 212) for ξ in slightly different situation with covariates.

For the second set of simulations, the data are simulated from a ZITPo distribution with parameters

$$\pi_i = \frac{\exp(\mathbf{x}_{i1}^T \boldsymbol{\beta}_1)}{1 + \exp(\mathbf{x}_{i1}^T \boldsymbol{\beta}_1)} \text{ and } \mu_i = \exp(\mathbf{x}_{i2}^T \boldsymbol{\beta}_2).$$

For the covariates, the first column of $\mathbf{X}$ is a column vector of 1 corresponding to the constant. The other columns of $\mathbf{X}$ were constructed with random values of respectively a normal, a Poisson, two binomials and an exponential distribution, with corresponding regression parameters $\boldsymbol{\beta}_1 = [1, 1, -0.5, 0.5, 0.25, 0.25]^T$ and $\boldsymbol{\beta}_2 = [2, 1, 0.5, 0.5, 0.25, 0.25]^T$. The values of the $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ were chosen in order to obtain asymmetrical distributions for the probabilities of positive outcomes, π_i, and for the expectations of positives values, μ_i. Figure 5 shows their respective distributions. With a median of 0.3, the probabilities of positive outcomes, π_i, are rather low. The expectations of the positives values, μ_i have a very asymmetrical distribution. The cutting value y° is a fixed value independent of i and which approximately corresponds to the quantile 0.1 of the positive simulated data. The choice of the parameter values π_i, μ_i and y° correspond to an extreme choice to test the performance in the MLE in non-trivial situations.

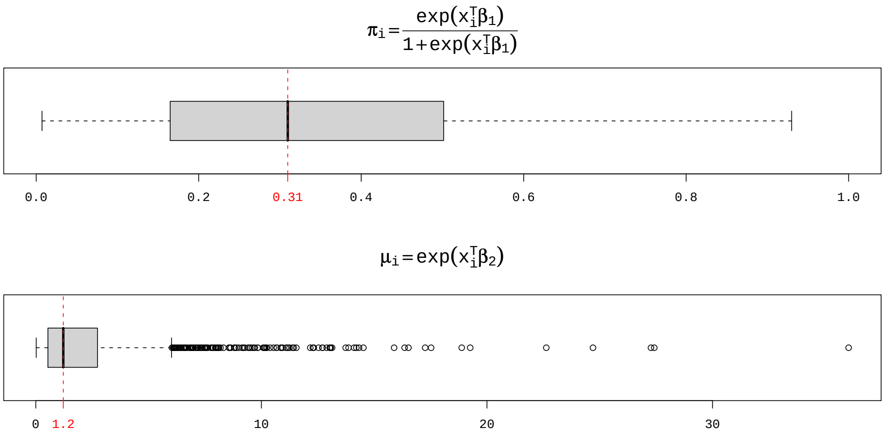

Figure 5: *Distribution of the probabilities of positive outcomes, π_i, and of the expectations of positives values, μ_i, used to simulate ZITPo realizations.*

The bottom plot of Figure 6 shows the sampling distribution of the MLE of the shape parameter ξ. The boxplots of the parameters estimates of ξ show an small underestimation of the parameter value even when the number of positive data is around 650 observations which correspond to 30% of the maximum sample size of this analysis. The $\xi-$estimates seem to converge more slowly when there are covariates. The bias of the shape parameter seems to depends on both the number of observations n and on the number of covariates p, a situation similar to the MLE of the parameter σ in multiple regression analyses.

The upper and centered plots of Figure 6 present the sampling distributions of the MLE of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. Regardless of the sample size, all boxplots are well centered around the true value of the parameters and the coverage levels of the corresponding confidence intervals are close to the 95% nominal value.

Note that $\boldsymbol{\beta}_2$ and the ξ are essentially estimated over the positive part of the data which represent the 30% of the 500, 1000 et 2000 observations of our study. Hence, our results appear very satisfactory. Similar results were obtained in simulations with $\xi = 0.5$.

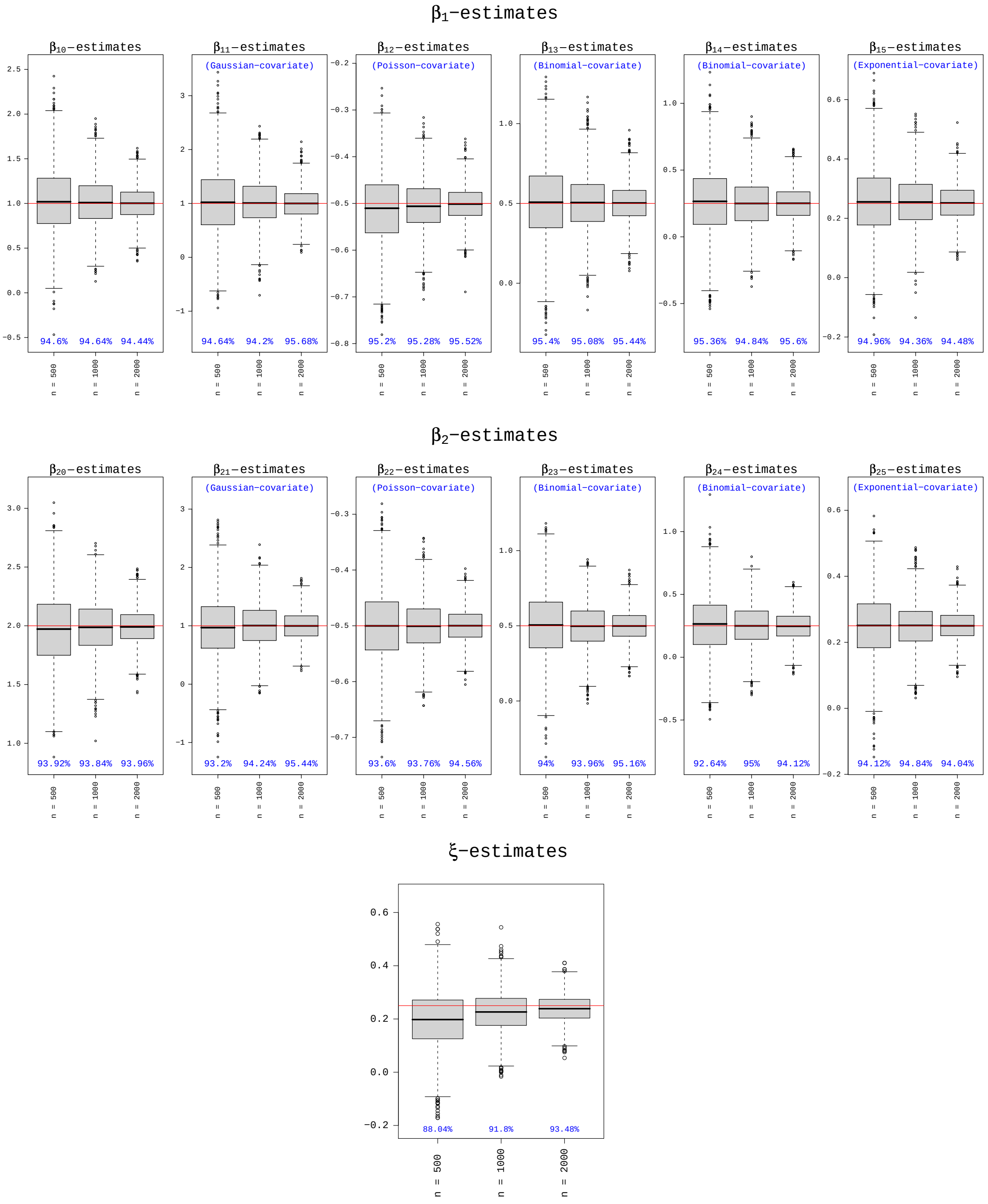

Figure 6: *Boxplots of the MLE of* $\boldsymbol{\beta}_1$ *(upper plots),* $\boldsymbol{\beta}_2$ *(centered plots) and* ξ *(bottom plot) computed over 2500 datasets simulated from a ZITPo distribution with parameters* $\boldsymbol{\beta}_1 = [1, 1, -0.5, 0.5, 0.25, 0.25]^T$, $\boldsymbol{\beta}_2 = [2, 1, 0.5, 0.5, 0.25, 0.25]^T$ *and* $\xi = 0.25$. y° *is a fixed value which approximately corresponds to the quantile 0.1 of the positive simulated data. Analyzes were performed for samples of sizes 500, 1000 and 2000. The horizontal red lines indicate the position of the true parameter values. The blue percentages indicate the coverage levels of confidence intervals of the form* $[\hat{\theta} - \Phi^{-1}(0.975)\hat{\sigma}_{\hat{\theta}}, \hat{\theta} + \Phi^{-1}(0.975)\hat{\sigma}_{\hat{\theta}}]$, *where* Φ *is the density function of the standard normal distribution.*

4 Model validation

Residual analyses in the context of zero-inflated models like described in (2) and (3) may be split in two parts: A first one focusing on the distribution that distinguishes the zeroes from the positive outcomes, and a second one considering the distribution of the positive values. In models with covariates, the residuals of the part distinguishing the zeroes correspond to residuals of logistic regressions. As this topic is already well covered in the literature (we refer to Collett (2003) for a complete overview), the following subsections focus on the residuals of the positive part of the model. We propose a residual type for truncated and untruncated generalized Pareto models. Examples based on simulated datasets with and without covariates are also presented.

Let $Y_i^{*+} = (Y_i^*|Y_i^* > 0)$ denote the positive values of the model and let $Y_i^+ = (Y_i|Y_i > y^\circ)$ be the observed truncated positive values. As $(Y_i^{*+} - y^\circ|Y_i^{*+} > y^\circ) = (Y_i^+ - y^\circ)$ and follows a GPD$(\mu_i + \frac{\xi y^\circ}{1-\xi}, \xi)$, let define the ith residual, ϵ_i, in the following way:

$$\epsilon_i = h(Y_i^+ - y^\circ) = \frac{Y_i^+ - y^\circ}{\mathrm{E}[Y_i^+ - y^\circ]} = \frac{Y_i^+ - y^\circ}{\mu_i + \frac{\xi y^\circ}{1-\xi}}. \tag{16}$$

The residuals distribution, $f_{\epsilon_i}(\epsilon_i)$, may then easily be derived and is given by

$$f_{\epsilon_i}(\epsilon_i) = f_{(Y_i^+ - y^\circ)}\left(h^{-1}(\epsilon_i)\right)\left|\frac{\partial}{\partial\epsilon_i}h^{-1}(\epsilon_i)\right| = \frac{1}{1-\xi}\left(1 + \frac{\xi}{1-\xi}\epsilon_i\right)^{-\frac{1}{\xi}-1}. \tag{17}$$

Thus $f_{\epsilon_i}(\epsilon_i) \sim$ GPD$(\mu = 1, \xi)$. The residuals theoretically (i.e. if the ZITPo model holds) follow a generalized Pareto distribution of parameters $\mu = 1$ and ξ. This result holds also when $y^\circ = 0$. Note that this result is not asymptotic and hence holds for any sample size, a pretty rare situation in GLM.

A very powerful and non asymptotic model validation procedure thus consists in comparing the distribution of the estimated residuals to their estimated theoretical distribution. The former are obtained by substituting in (16) the parameters by their estimated values, i.e.

$$\hat{\epsilon}_i = \frac{Y_i^+ - y^\circ}{\hat{\mu}_i + \frac{\hat{\xi} y^\circ}{1-\hat{\xi}}} \sim \text{GPD}(\mu = 1, \hat{\xi}),$$

QQ-plots should approximately display a straight line when the model adequately fits the data.

Finally, the result in (17) offers a fast method to generate random realizations from truncated or untruncated generalized Pareto models. Indeed, let $\tilde{\mathbf{u}}$ be a vector of n random realizations of a Uniform(0,1) and let $\boldsymbol{\mu} = [\mu_1, ..., \mu_n]^T$ be the vector of expectations of the generalized Pareto distribution. Then, inverting (16) and (17) allows to generate $\mathbf{y}$, a vector of n random variates of a y°-truncated GPD$(\boldsymbol{\mu}, \xi)$, in the following way:

$$\mathbf{y} = \left[\left(\tilde{\mathbf{u}}^{-\xi} - 1 \right) \frac{1-\xi}{\xi} \right] \left(\boldsymbol{\mu} + \frac{\xi y^\circ}{1-\xi} \right) + y^\circ.$$

The residuals of the generalized Pareto models with covariates studied the previous section are analyzed for one simulated dataset. The aim is to check if theoretical behavior of the residuals is suitable in real samples. As the quantiles of the generalized Pareto distribution are very asymmetrical, QQplots of the log of the residuals are also proposed.

Figure 7 presents the residual plots of the analysis of a dataset of 1000 observations simulated from a ZITPo distribution incorporating covariates. The parameter values are $\boldsymbol{\beta}_1 = [1, 1, -0.5, 0.5, 0.25, 0.25]^T$, $\boldsymbol{\beta}_2 = [2, 1, 0.5, 0.5, 0.25, 0.25]^T$, $\xi = 0.25$ and y° corresponds to the quantile 0.1 of the positive data.

The ordered residuals are compared to the quantiles of the estimated theoretical distribution, a GPD$(1, \hat{\xi} = 0.274)$. The QQplots of the residuals and of their log show a good adequacy of the model to the data.

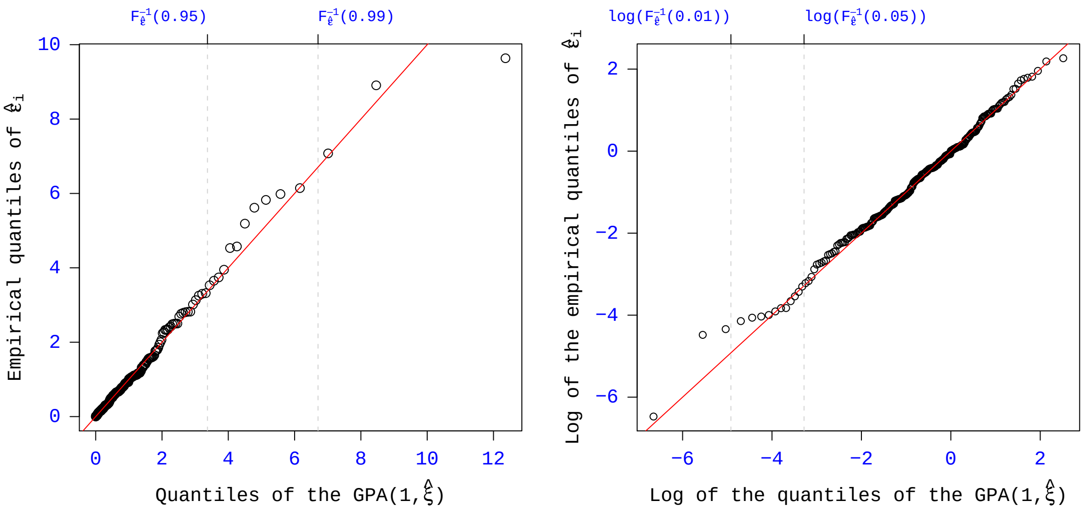

Figure 7: *QQplots of the residuals (left) and of the log of the residuals (right) of a dataset of 1000 observations simulated from a ZITPo distribution with covariates. The ordered residuals are compared to the quantiles (left) and to the log of the quantiles (right) of a* GPD$(1, \hat{\xi} = 0.274)$.

5 Applications to radio audience data

The ZITPo model is applied to the audience data of the local radio station “116” in its broadcasting area during the weekdays of the second semester of 2007. The upper plot of Figure 8 presents the distribution of the daily listening times of 2155 participants measured during one day of this period. The clump-at-zero represents 63% of the data.

The audience indicators of rating and time spent listening are explained by a set of categorical variables including the age in 5 classes ([15-25[, [25-35[,[35-45[,[45-60[,[60-120[), the education level in 3 classes (low, mid, high), the gender, the time in month and the different zones of the broadcast area. The contrasts used to create the $k-1$ dummy variables from a k-classes categorical variable are of type “treatment” for the variables age, gender and education with base “15 to 25 years old men with low education level”, and of type “Sum” for the geographical zones and the months. The model includes interaction between age and gender. Other interactions – like between education and age – appeared non-significant and did not improve the log-likelihood or the residual distribution.

To protect the parameter estimates of the possible influence of the false positive and false zeroes observations belonging to the interval $[0, 5[$, we choose $y^{\circ} = 4.95$. Consequently, we coded the 15 observations belonging to the interval $[3, 5[$ in Figure 8 as zeroes and let the ZITPo model adequately separate the true from the false zeroes as described in the first part of (10).

The $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ estimated values as well as their standard deviations are reported in Table 1. The p-values corresponding to the (asymptotic) significance tests for $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, i.e. $2\Phi^{-1}(-|\hat{\beta}/\hat{\sigma}_{\hat{\beta}}|)$ are also indicated. According to the chosen contrasts, the estimated intercepts β_{10} and β_{20} are related to the estimated rating and time spent listening of 15 to 25 years old men with a low education level through respectively

$$\frac{\exp(\hat{\beta}_{10})}{1+\exp(\hat{\beta}_{10})} \cong 0.12 \text{ and } \exp(\hat{\beta}_{20}) \cong 59.$$

15 to 25 years old men living in the broadcast area of interest and having a low education level have thus a probability of contact to radio station "116" of 12% and an average contact length of about 59 minutes during the second semester of 2007. The estimated distribution of the effective (untruncated) positive times of the individuals of this focus group is thus:

$$Y_i^*|(Y_i^* > 0) \sim \text{GPD}(59, \hat{\xi} \cong 0.08).$$

Thus, under the model, $F^{-1}_{(Y^*|Y^*>0)}(3|59,\hat{\xi}) \cong 0.05$ and $F^{-1}_{(Y^*|Y^*>0)}(y^\circ|59,\hat{\xi}) \cong 0.09$ respectively represent for this focus group the estimation of the part of effective positive data that is coded as zero by the swiss measurement system and the estimation of the part of the effective positive data that was supposed truncated and coded as zero for the estimation. The average ratings and time spent listening of other focus groups – like women with a high education level – are then shifts of 12% and 59 minutes.

In order to test the significance of each factor (e.g. age), we use the likelihood ratio test to compare nested models. Let $\boldsymbol{\beta} = [\boldsymbol{\beta}^T_{(1)}, \boldsymbol{\beta}^T_{(2)}]^T$ be the vector of the regression parameters. The LRT statistic can be used to test hypotheses of the form $H_0 : \boldsymbol{\beta}^T_{(2)} = 0$ against $H_1 : \boldsymbol{\beta}^T_{(2)} \neq 0$ (with $\boldsymbol{\beta}^T_{(1)}$ unspecified) and is given by

$$LRT = 2\big[l(\hat{\boldsymbol{\beta}}|\mathbf{y}, y^\circ, \mathbf{X}_1, \mathbf{X}_2) - l(\dot{\boldsymbol{\beta}}|\mathbf{y}, y^\circ, \mathbf{X}_1, \mathbf{X}_2)\big],$$

where $\hat{\boldsymbol{\beta}}$ and $\dot{\boldsymbol{\beta}}$ respectively denote the full and reduced regression parameters

MLE. The LRT statistic follows a $\chi^2_{p-\dot{p}}$ distribution under the null hypothesis, where p and $\dot{p}$ are the number of parameters of the full and reduced model.

Table 2 presents the LRT evaluating which variables significantly influence the rating and the average listening times. According to the corresponding p-values, the variables significantly influencing the average rating are the age, the education level and the geographical zone in the broadcast area. A look at the $\boldsymbol{\beta}_1$ estimates shows that the rating average increases with age and education classes and decreases for people living in the countryside area named "Zone 2". The variables significantly influencing the average listening time are the age, the gender and area. The listening time average increases for people belonging to high age classes and decreases for people living in "Zone 2". The evolution of listening time with age is not the same for men and women.

The estimated shape parameter is $\hat{\xi} = 0.082$ with $\hat{\sigma}_{\hat{\xi}} = 0.039$. The shape parameter is thus slightly but significantly higher than zero. The residuals are to be compared to a $\text{GPD}(1, 0.082)$. The analysis of the fit is presented in the two bottom plots of Figure 8. The QQplots of the residuals and of their log show a very good adequacy of the model to the data.

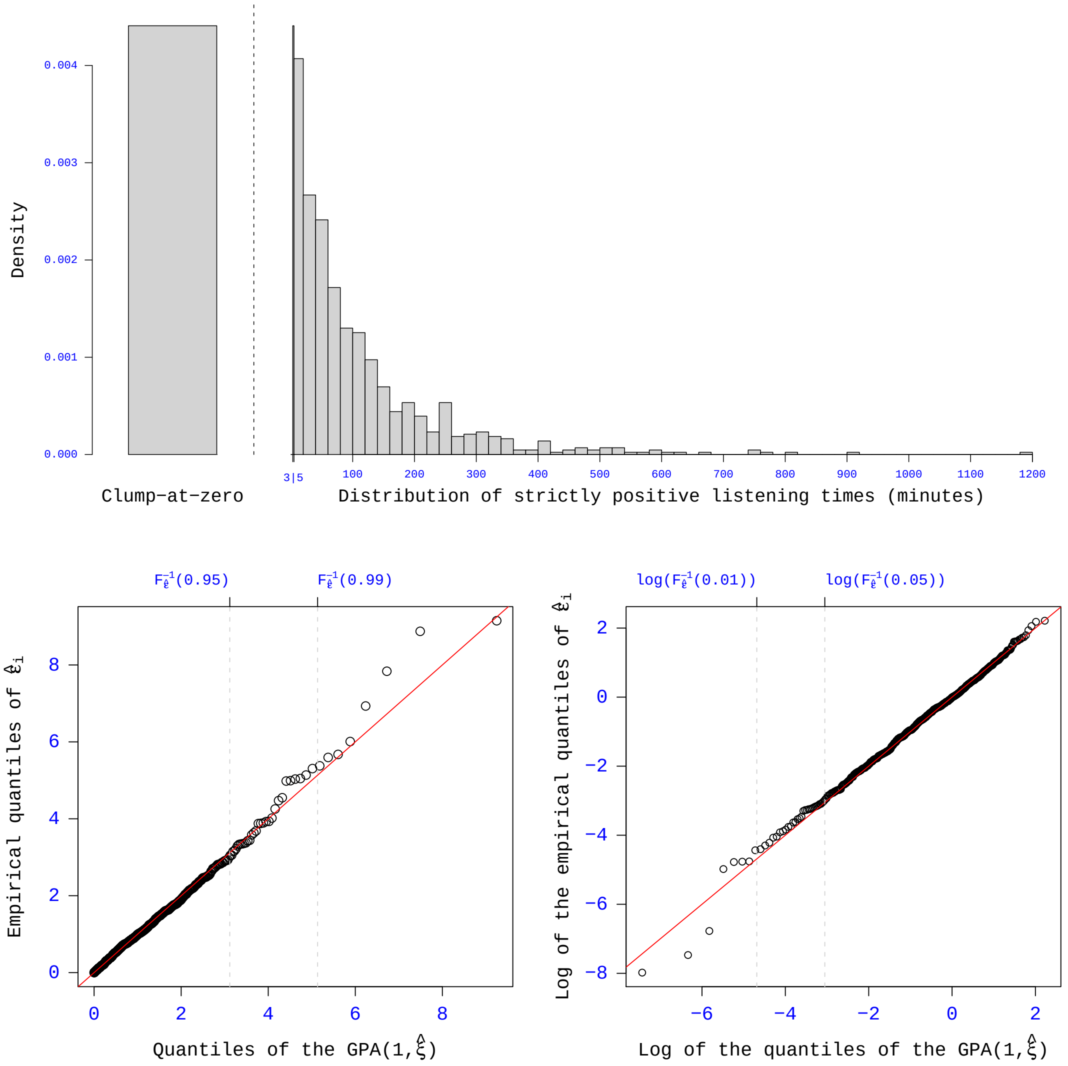

Figure 8: *Upper plot: Distribution of the observed listening times to radio "116" during the second semester of 2007. The number of observations is 2155. The clump-at-zero represents the 63% of the data. Two bottom plots: QQplots of the residuals (left) and of the log of the residuals (right) of the ZITPo model applied to the listening times to radio "116". The ordered residuals are compared to the quantiles (left) and to the log of the quantiles (right) of a* $\mathrm{GPD}(1, \hat{\xi} = 0.082)$.

	Rating				Average Listening Time			
	$\hat{\boldsymbol{\beta}}_1$	SE	p-value	Sig.	$\hat{\boldsymbol{\beta}}_2$	SE	p-value	Sig.
(Intercept)	−1.95	0.32	<0.001	***	4.08	0.33	<0.001	***
[25 − 35[	0.40	0.39	0.309		−0.16	0.39	0.680	
[35 − 45[	0.94	0.36	0.008	**	0.20	0.35	0.568	
[45 − 60[	1.57	0.34	<0.001	***	0.40	0.34	0.235	
[60 − 120[	2.22	0.35	<0.001	***	0.76	0.34	0.026	*
Women	−0.25	0.49	0.608		−0.73	0.49	0.133	
Educ. Middle	0.18	0.16	0.255		0.01	0.13	0.933	
Educ. High	0.36	0.12	0.002	**	−0.15	0.09	0.103	
July	−0.15	0.12	0.216		−0.06	0.10	0.516	
August	−0.11	0.11	0.346		0.04	0.09	0.695	
September	0.14	0.11	0.225		−0.07	0.09	0.393	
October	0.18	0.11	0.085	.	0.01	0.08	0.948	
November	−0.00	0.11	0.973		0.04	0.09	0.654	
Zone 2	−0.26	0.05	<0.001	***	−0.08	0.04	0.049	*
Women +[25 − 35[	−0.02	0.58	0.970		1.26	0.57	0.028	*
Women +[35 − 45[	0.18	0.54	0.737		0.71	0.53	0.180	
Women +[45 − 60[	0.03	0.52	0.961		0.90	0.51	0.079	.
Women +[60 − 120[	0.39	0.52	0.455		1.11	0.50	0.027	*

Table 1: *$\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ estimated parameters and corresponding standard deviations of the ZITPo model applied to the listening times to radio station "116" . The p-values are for (asymptotic) significance testing of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. Low p-values are magnified in the columns "Sig." by means of (***), (**), (*), (.) respectively corresponding to significant tests at the levels of 0.001, 0.01, 0.05 and 0.1.*

	Rating				Average Listening Time			
	T	Df	p-value	Sig.	T	Df	p-value	Sig.
Age + Age·Gender	236.58	8	<0.001	***	78.54	8	<0.001	***
Gender + Age·Gender	3.26	5	0.659		16.08	5	0.007	**
Education	9.61	2	0.008	**	3.41	2	0.182	
Month	5.91	5	0.315		1.53	5	0.909	
Zone	24.67	1	<0.001	***	3.92	1	0.048	*
Age·Gender	2.56	4	0.634		8.14	4	0.087	.

Table 2: *LRT statistics (with corresponding degrees of freedom) and p-values for the marginal LRT applied to the listening times to radio station "116". Each variable (or variable plus interaction) of the left column are tested in the binomial (Rating) and truncated GPD (Average Listening Time) part of the model. Low p-values are magnified in the columns "Sig." by means of (***), (**), (*), (.) respectively corresponding to significant tests at the levels of 0.001, 0.01, 0.05 and 0.1.*

6 Conclusion

The ZITPo model is a very powerful model that can be used in particular to analyze radio audience data. Using the truncated observations, this model allows to adequately estimate the true proportions of non-zero observations and the average of positive values – corresponding to the audience indicators of rating and time spent listening – of the underlying untruncated listening times distribution. The model also allows to relate these expectations to covariates in a GLM spirit, providing an explanatory model to audience data. The model validation procedure resulting from properties of the generalized Pareto distribution offers a very helpful way to judge the adequacy of the model to the data.

Although the main motivation for the development of the ZITPo model was the analysis of radio audience data, we believe that it can adequately fit a number of datasets which have heavy tails distributions. For example it provides an extension to model (4) for hydrological data, that can include covariates to explain the mean level, with $y^\circ = 0$.

Acknowledgments

This work was supported by the Swiss National Science Foundation. Radio data, measured by the Radiocontrol measurement system, were kindly provided by the Mediapulse Corporation.

References

Aitkin, M. and D. Clayton (1980). The fitting of exponential, Weibull and extreme value distributions to complex censored survival data using GLIM. *Applied Statistics 29*, 156–163.

Beirlant, J., P. Vynckier, and J. L. Teugels (1996). Tail index estimation, Pareto quantile plots, and regression diagnostics. *Journal of the American Statistical Association 91*, 1659–1667.

Castillo, E. and A. S. Hadi (1997). Fitting the generalized Pareto distribution to data. *Journal of the American Statistical Association 92*, 1609–1620.

Chapados, N., Y. Bengio, V. Vincent, J. Ghosn, C. Dugan, I. Takeuchi, and L. Meng (2002). Estimating car insurance premia: A case study in high-dimensional data inference. In *Advances in Neural Information Processing*, Volume 14, pp. 1369–1376.

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

Chen, Y., Y. Jiang, and Y. Mao (2007). Hospital admissions associated with body mass index in canadian adults. *International Journal of Obesity 31*, 962 – 967.

Christmann, A. (2004). An approach to model complex high-dimensional insurance data. *Allgemeines Statistisches Archiv 88*, 375–397.

Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values.* Springer-Verlag Inc.

Collett, D. (2003). *Modelling Binary Data.* Chapman & Hall Ltd.

Davison, A. C. and R. L. Smith (1990). Models for exceedances over high thresholds (with comments). *Journal of the Royal Statistical Society, Series B: Methodological 52*, 393–442.

Duan, N., J. Manning, Willard G., C. N. Morris, and J. P. Newhouse (1983). A comparison of alternative models for the demand for medical care. *Journal of Business & Economic Statistics 1*, 115–126.

Dupuis, D. J. and M. Tsao (1998). A hybrid estimator for generalized pareto and extreme-value distributions. *Communications in Statistics: Theory and Methods 27*(4), 925–941.

Dupuis, D. J. and M.-P. Victoria-Feser (2006). A robust prediction error criterion for Pareto modelling of upper tails. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique 34*(4), 639–658.

Hall, P. and A. H. Welsh (1985). Adaptive estimates of parameters of regular variation. *The Annals of Statistics 13*, 330–341.

Heindervckx, F. and A. Phillips (2001). Mesurer les audiences à l'époque de la convergence médiatique. In *Enquête, modèles et applications*, Chapter 5.1, pp. 231–241. Dunod.

Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. *The Annals of Statistics 3*, 1163–1174.

Hosking, J. R. M. and J. R. Wallis (1987). Parameter and quantile estimation for the generalized Pareto distribution. *Technometrics 29*, 339–349.

Juárez, S. F. and W. R. Schucany (2004). Robust and efficient estimation for the generalized Pareto distribution. *Extremes 7*(3), 237–251.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics 34*, 1–14.

Min, Y. and A. Agresti (2002). Modeling nonnegative data with clumping at zero: A survey. *Journal of the Iranian Statistical Society 1*(1-2), 7–33.

Min, Y. and A. Agresti (2005). Random effect models for repeated measures of zero-inflated count data. *Statistical Modelling 5*(1), 1–19.

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics 33*, 341–365.

Nelder, J. A. and R. W. M. Wedderburn (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A: General 135*, 370–384.

Peng, L. and A. H. Welsh (2001). Robust estimation of the generalized Pareto distribution. *Extremes 4*(1), 53–65.

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

Ridout, M., C. G. Demétrio, and J. Hinde (1998). Models for count data with many zeros. In *International Biometric Conference*, pp. 179–192.

Singh, V. P. and M. Ahmad (2004). A comparative evaluation of the estimators of the three-parameter generalized Pareto distribution. *Journal of Statistical Computation and Simulation 74*(2), 91–106.

Weglarczyk, S., W. G. Strupczewski, and V. P. Singh (2005). Three-parameter discontinuous distributions for hydrological samples with zero values. *Hydrological Processes 19*(15), 2899–2914.

Welsh, A. H., R. B. Cunningham, C. F. Donnelly, and D. B. Lindenmayer (1996). Modelling the abundance of rare species: statistical models for counts with extra zeros. *Ecological Modelling 88*, 297–308.