Inference with extremes: Accounting for Extreme Values in Count Regression Models

David Randahl[1] and Johan Vegelius[2]

[1] *Department of Peace and Conflict Research, Uppsala University, Email: david.randahl@pcr.uu.se*
[2] *Department of Statistics, Uppsala University, Email: johan.vegelius@statistik.uu.se*

Abstract

Processes which occasionally, but not always, produce extreme values are notoriously difficult to model as a small number of extreme observations may have a large impact on the results. Existing methods for handling extreme values are often arbitrary and leave researchers without guidance regarding this problem. In this paper we propose an Extreme Value and Zero Inflated Negative Binomial (EVZINB) regression model, which allows for separate modeling of extreme and non-extreme observations, to solve this problem. The EVZINB model offers an elegant solution to modeling data with extreme values and allows researchers to draw additional inferences about both extreme and non-extreme observations. We illustrate the usefulness of the EVZINB model by replicating a study on one-sided violence against civilians.

1 Introduction

As the availability and resolution of data increases in the political and social sciences, researchers are increasingly faced with difficult modeling decisions when the data they are trying to model do not conform to well-behaved and easily modelled distributions. For example, when modeling the number of fatalities from one-sided violence (OSV) against civilians (Eck and Hultman, 2007) on a country-month or country-year level it can be assumed that certain countries will never experience any fatalities from OSV simply because they are not at risk of political violence and thus produce ’structural’ zeroes (see for instance Hultman, 2013; Bagozzi, 2015). Conversely, in other countries there may be periods of very large counts of OSV during campaigns of ethnic cleansing or genocide, which can be thought of as a separate process from the vast majority of cases. Fitting a regular count regression model in these circumstances may produce biased parameter results if the factors which cause the structural zeroes or the extreme counts are different from the factors which cause the non-extreme positive counts. The former of these problems, that of excessive, structural, zeroes, has received much attention in the literature and it has been shown that zero-inflated count models are appropriate from both an empirical and theoretical level, as they allow for more stable estimation, more accurate predictions, and more nuanced inferences (for a longer discussion on zero-inflated models see Greene, 1994; Bagozzi, 2015).

The latter problem, that of the influence of extreme values on the estimation of count regression model, has, however, to our knowledge received little or no attention in the literature, and it is this issue that this paper aims to address. In this paper we propose an *extreme value and zero inflated negative binomial* (EVZINB) regression model to allow for modeling data with extreme values. This regression model allows the researcher to model data which contains extreme values without arbitrary inclusion or exclusion criteria, and allows the researcher to draw inferences about which factors influence the likelihood of extreme values and which factors influence the ’extremeness’ of the extreme values. These types of inferences are especially useful for processes where certain factors may have limited effect on the average cases but large effects on the more extreme cases. In these processes it may be especially relevant to identify these ’risk-factors’ which may cause the process to exhibit extreme value behavior. In addition to showing that the EVZINB model allows researchers to draw new and more nuanced inferences, we also show that failing to account for extreme values in the analysis risks causing biased parameter estimates.

This paper proceeds by motivating the EVZINB model empirically, in section 2, and statistically, in section 3. We also employ a simulation study to show that the EVZINB model manages to obtain accurate parameter estimates in data generated with excessive zeroes and extreme values, while the regular negative binomial

(NB) and zero-inflated negative binomial (ZINB) models produce biased estimates under those circumstances. We then show the empirical utility of the EVZINB model by replicating a study on one-sided violence against civilians by rebel groups (Moore, 2019). In this replication we show that using the EVZINB model we are able to draw more fine-grained inferences on how different factors shape the OSV rebel groups, and that the EVZINB model outperforms the NB and ZINB models on a number of crucial evaluation metrics. In the final section the implications of the introduction of the EVZINB model are discussed, and future potential extensions explored.

2 Motivating the Extreme Value Inflated model

Extreme values, or extreme observations, tend to arise in a multitude of different disciplines where the phenomena studied have a self-reinforcing component. However, extreme values need not only appear from processes which often or always follow an extreme value distribution. Rather, extreme values can also appear in phenomena which in the majority of cases produce values from non-extreme distributions. Examples of such phenomena include different types of organized political violence (for instance Lacina and Gleditsch, 2005; Hultman, 2013; Eck and Hultman, 2007; Clauset, Young, and Gleditsch, 2007), crime rates (Disha, 2019), and mass protests (Weidmann and Rød, 2019), but could just as well be applied to a multitude of other topics such as the number of deaths in a pandemic, the number of followers on social media, or analyses of number of closed trades in stocks. Outside of the social sciences, similar phenomena have been observed in widely different fields like the sizes of solar flares (Litvinenko, 1996), the use of word distribution in languages (Cancho and Solé, 2003) and the earthquake sizes in California (Gutenberg and Richter, 1944)

The fact that extreme values exist but are rare do, however, cause a number of different problems, theoretical as well as empirical or methodological, for researchers aiming to model the phenomena quantitatively. On the theoretical level, it may seem odd to include the most extreme observations of the phenomenon of interest in the analysis as it may well be argued that the observation arises from a different process. For instance, when studying one-sided violence against civilians (OSV), a researcher may well argue that cases where an active genocide is ongoing should be excluded as genocide arise from a different process than other forms of OSV. On the other hand, excluding the most prominent cases of OSV may also seem like a strange choice for this researcher. Worse yet, since the extreme values by their nature are very large compared to the vast majority of cases the extreme values tend to have a large effect on the results of any type of quantitative modeling, meaning that the decision to include or exclude certain cases may severely affect the results of the modeling. To not be forced to make an arbitrary inclusion or exclusion decision for a single or handful of case(s), the researcher may decide to try one of a multitude of ’objective’ solutions, such as ’trimming’ (excluding) or ’winsorizing’ (censoring observations to a threshold value) (Dixon and Yuen, 1974) a certain number or percentage of cases. Yet, these techniques are in most cases neither statistically nor theoretically sound as they simply mask the arbitrary nature of the inclusion or exclusion criteria, and may severely bias the results by either removing important observations or artificially changing values on the dependent variable.

To show the effects of the inclusion, exclusion, or censoring of extreme values we created a simple example regression model, where we modelled the country-month counts of OSV in Africa between 1989 and 2019 (Pettersson and Öberg, 2020) against the natural logarithm of the population and two dummy variables indicating democracy and autocracy (Hegre et al., 2019). We modeled this using both a regular negative binomial (NB) regression model, and a zero-inflated negative binomial (ZINB) regression model. In the *original* models we used all available country-month observations, in the *trimmed* models we removed the 10 largest counts, and in the *winsorized* models we censored the 10 largest observations to the 11th largest value in the dataset. The results of these regressions can be found in 1 below.

Table 1. Regression results for simple model of OSV counts with different modes of handling extreme values

	NB original	NB trimmed	NB winsorized	ZINB original	ZINB trimmed	ZINB winsorized
log(pop)	0.038***	0.035***	0.035***	0.001	0.011***	0.013***
	(0.003)	(0.003)	(0.003)	(0.004)	(0.003)	(0.003)
autocracy	−1.536***	0.305**	0.245**	−0.673***	0.710***	0.642***
	(0.135)	(0.119)	(0.120)	(0.177)	(0.125)	(0.128)
democracy	−4.251***	−2.063***	−2.254***	−3.583***	−1.516***	−1.831***
	(0.110)	(0.097)	(0.098)	(0.128)	(0.105)	(0.104)
Constant	3.811***	1.678***	1.876***	4.228***	2.100***	2.228***
	(0.079)	(0.069)	(0.070)	(0.073)	(0.065)	(0.065)
Observations	19,378	19,368	19,378	19,378	19,368	19,378
θ	0.022	0.029	0.028	0.028	0.039	0.036
AIC.	39,193	37,786	38,147	38,079	36,680	36,984

Note: *p<0.1; **p<0.05; ***p<0.01

The results in the table above show that by trimming or winsorizing the 10 largest counts, i.e. the 0.05% most extreme values, in our dataset the coefficient for autocracy goes from being negative and highly statistically significant to being positive and highly statistically significant. Relatedly, the coefficient for democracy, while staying in the same direction and level of significance, is roughly halved for both the NB and ZINB specifications. This example may be simplistic in terms of the covariates included in the model, but it clearly highlights the effects of the researchers' choice of including or excluding certain observations in the analysis. Without clear theoretical guidance, the researcher is left to decide inclusion or exclusion criteria based on perhaps arbitrary decision rules which may ultimately severely affect the results of any study.

The problems with extreme values are not limited to their effect on the observed results, it may also affect the estimation method and the possibility to make diagnostic tests or alternative specifications of the models. Depending on what regression model is used for the analysis, it may or may not converge depending on whether or not certain extreme values are included or not. When re-running the NB regression model from Table 1 using bootstrapping, the algorithm failed to converge in approximately 1.7% of the bootstraps. Changing the covariate log(population) to log(gdp/capita) causes the estimation to fail in approximately 9.4% of cases. When estimating the same models without the 10 largest counts, the algorithm does not fail in any of the bootstrapped cases, regardless of whether log(population) or log(gdp/capita) is used. This shows that not only may failing to properly deal with the extreme values cause biased results, it may also make certain tools of analysis unavailable to the researcher.

2.1 The EVZINB-model

To alleviate the problems associated with modeling processes which sometimes but not always exhibit extreme values we propose the extreme value and zero-inflated negative binomial (EVZINB) regression model[1], a novel extension of the zero-inflated negative binomial regression model. This model deals with the problems caused by extreme values in count data by including a separate component for modeling these extreme values. The inspiration for the EVZINB model stems from zero-inflated regression models(Greene, 1994), where data are assumed to originate from two separate processes; one process generating zeroes, and one count process (which may also produce zeroes). In the zero-inflated regression models, these two processes are modelled separately and can be thought of as two different components of the models which may include different covariates. The first component aims to model structural, excess, zeroes using one set of covariates, and the second component models the count process separately from these 'excess zeroes'. Our proposal is to extend this framework to a three component regression model, where both excess zeroes and extreme values are modelled separately. This three component regression model can be seen as a regression model with latent states, where the latent states represent the sub-processes from which the data are generated.

The benefits from this approach are manifold. First, by allowing for three separate states in the model, each of the states can be estimated while filtering out the effects of the other two. This will lead to more stable and less biased parameter estimates for each of the processes, and to fewer convergence issues in large data. Secondly, it allows the researcher to specify different covariates on each of the processes, allowing for a more nuanced analysis. Third, filtering out the effects of extreme observations should give more stable out of sample predicted values from the model, enhancing the predictive performance of the estimated models. Fourth, the regression model allows for the estimation of the probability of any given observation being part of either of the three states, widening the possibilities for analysis further.

Aiming to model extreme values is nothing new. Multiple studies within the conflict sciences use extreme value analysis to discuss grand questions such as the overall decline in large deadly wars, the distribution of mass atrocities, or terrorist attacks (see for instance Clauset, 2017; Clauset, Young, and Gleditsch, 2007; Cunen, Hjort, and Nygård, 2020; Cirillo and Taleb, 2016). Common for all of these approaches is that the researchers are explicitly interested in drawing inferences about the tail risks and the shapes of the extreme value distributions. However, to our knowledge no attempts have been made at integrating such approaches to provide a unified framework to analyze both the extreme and non-extreme states of a process simultaneously. While the EVZINB model allows the researcher to draw inferences about the extreme value state of the model, it is not necessary for the researcher to do so. For example, by choosing not to include any covariates on the extreme value state, the EVZINB model would simply work as a filter to filter out the excess effect of extreme values in the analysis, just as a zero-inflated model without any covariates on the zero-inflation state filters out excessive zeros without giving any information as to what causes these excessive zeroes.

3 Statistical Methodology

In the presence of count data a common approach is to employ the Poisson regression models (e.g., Frome, Kutner, and Beauchamp, 1973). However, the Poisson distribution is associated with the mean-variance equality restriction, conditional on explanatory variables. It has been recognized, (e.g. Dean and Lawless, 1989) that count data commonly display overdispersion, i.e., the variance is larger than the mean. One possibility to account for overdispersion is the negative binomial (NB) regression, which relaxes the mean-variance equality restriction (for example Lawless, 1987). It is, however, also common with data exhibiting a proportion of excess zeroes which cannot be appropriately modeled using the NB regression model. The zero-inflated negative binomial (ZINB) regression model accounts for excess zeroes by introducing a proportion of zeroes which can be modeled using logistic regression. See for example (Minami et al., 2007; Greene, 1994).

[1] This model can also be use when zero inflation is not present in which in which case it would reduce to an extreme value inflated negative binomial regression model, EVINB.

The latent states associated with only zeros and moderate count data will denoted Z and NB, respectively.

In addition to excess zeroes, it is also common to observe a substantial proportion of extremely large values leading to unstable estimates or non-converging models using NB or ZINB. We propose to introduce another latent state, generating extreme values, alongside the extra latent states generating zeroes (Z) and moderate count data (NB). This state will be referred to as the EV state. Define the unobserved random variable $\boldsymbol{W}$ which is $\boldsymbol{W} = (1,0,0)^T$, $\boldsymbol{W} = (0,1,0)^T$ or $\boldsymbol{W} = (0,0,1)^T$ if the latent state is Z, NB or EV, respectively. Then we introduce the observed random variable Y with the following conditional properties:

$$\begin{aligned} Y|\boldsymbol{W} &= (1,0,0)^T = 0 \\ Y|\boldsymbol{W} &= (0,1,0)^T \sim \mathrm{NegBin}\left(\mu_{NB}, \alpha_{NB}\right) \\ Y|\boldsymbol{W} &= (0,0,1)^T \sim \mathrm{Pareto}\left(c_{EV}, \alpha_{EV}\right) \end{aligned}$$

where μ_{NB} and α_{NB} are the mean and dispersion parameters of a negative binomial distribution, respectively, α_{EV} and c_{EV} are the shape and cut-off parameters of a Pareto distribution and the prior probabilities $\pi_Z = \Pr\left(W = (1,0,0)^T\right)$, $\pi_{EV} = \Pr\left(W = (0,0,1)^T\right)$ and $\pi_{NB} = 1 - \pi_Z - \pi_{EV}$. The exact expressions for the probability mass functions of the negative binomial and Pareto distributions are given in the appendix. The distribution of the random variable Y defining the data-generating process in this study can be summarized as

$$Y \sim \mathrm{EVZINB}\left(\pi_Z, \pi_{EV}, \mu_{NB}, \alpha_{NB}, \alpha_{EV}, c_{EV}\right), \tag{1}$$

The prior probabilities (π_Z and π_{EV}) of the latent states are modeled using multinomial logistic regression and the extreme-value state is assumed to follow a Pareto distribution with shape parameter α_{EV} modeled as a function of the covariates. α_{EV} is a real, positive number governing the tail of the Pareto distribution. Whereas the Pareto distribution is proportional to $y^{-\alpha_{EV}}$, the exponential and Gaussian distributions are proportional to e^{-ky} and e^{-cy^2}, respectively, for positive numbers c and k. Hence, the Pareto distribution is associated with substantially more probability in its tail, and is hence, suitable to model extremely large observations. In fact, when α_{EV} approaches 1 from above, the expected value approaches infinity. Another property is that the conditional probability of yielding an observation twice the magnitude of some number $c_1 > c_{EV}$ is the same for any c_1. In this sense the Pareto distribution is scale invariant (Mandelbrot, 1983) and, thus, suitable for modeling self-enhancing phenomena like those discussed in the introduction. As described below, each observation will be associated with a shape-parameter α_{EV} which indicates the potential extremeness of a distribution with corresponding covariates. The full model will be referred to as an extreme-value and zero-inflated negative binomial (EVZINB) regression model. The model-implied latent-state probabilities, or the prior probabilities $\pi_{Z,i}$, $\pi_{NB,i}$ and $\pi_{EV,i}$ of observation i, are modeled as

$$\begin{aligned} \pi_{Z,i} &= \frac{\exp\left\{\boldsymbol{\gamma}_Z^T \boldsymbol{x}_{\pi,i}\right\}}{1 + \exp\left\{\boldsymbol{\gamma}_Z^T \boldsymbol{x}_{\pi,i}\right\} + \exp\left\{\boldsymbol{\gamma}_{EV}^T \boldsymbol{x}_{\pi,i}\right\}} \\ \pi_{EV,i} &= \frac{\exp\left\{\boldsymbol{\gamma}_{EV}^T \boldsymbol{x}_{\pi,i}\right\}}{1 + \exp\left\{\boldsymbol{\gamma}_Z^T \boldsymbol{x}_{\pi,i}\right\} + \exp\left\{\boldsymbol{\gamma}_{EV}^T \boldsymbol{x}_{\pi,i}\right\}}, \end{aligned} \tag{2}$$

where $\pi_{NB,i} = 1 - \pi_{Z,i} - \pi_{EV,i}$ and $\boldsymbol{x}_{\pi,i}$ is a column vector of covariates (starting with 1 for intercept)[2], and $\boldsymbol{\gamma}_Z$ and $\boldsymbol{\gamma}_{EV}$ are zero-inflation and extreme-value-inflation parameter vectors, respectively. The conditional mean $\mu_{NB,i}$ of the NB latent state of observation i, and the observation-specific shape $\alpha_{EV,i}$ of the EV latent

[2] It is possible to use different covariates for the Z and EV latent states but the same are used in the presentation for notational convenience.

state are modeled as

$$\begin{aligned} \mu_{NB,i} &= \exp\left\{\boldsymbol{\beta}_{NB}^T \boldsymbol{x}_{NB,i}\right\} \\ \alpha_{EV,i} &= \exp\left\{\boldsymbol{\beta}_{EV} \boldsymbol{x}_{EV,i}\right\}, \end{aligned} \tag{3}$$

where $\boldsymbol{x}_{NB,i}$ and $\boldsymbol{x}_{EV,i}$ are covariates (starting with 1 for intercept) and $\boldsymbol{\beta}_{NB}$ and $\boldsymbol{\beta}_{EV}$ are column vectors of parameters. Additional model parameters are the dispersion parameter of the NB latent state α_{NB} and c_{EV} which is the lower bound of observations from the Pareto distribution (EV latent state)[3]. The model parameters, hence, include $\boldsymbol{\gamma}_Z, \boldsymbol{\gamma}_{EV}, \boldsymbol{\beta}_{NB}, \boldsymbol{\beta}_{EV}, \alpha_{NB}$, and c_{EV}. Those are estimated with maximum likelihood using a version of the EM algorithm of Dempster, Laird, and Rubin, 1977, combining a generalized expectation maximization (GEM) algorithm, e.g. Wu, 1983; Lange, 1995, and an expectation conditional maximization either (ECME) algorithm of Liu and Rubin, 1994. The model estimation automatically provides the model-implied (ex-ante) latent state probabilities $\pi_{Z,i}$, $\pi_{NB,i}$ and $\pi_{EV,i}$ and the ex-post latent state probabilities after observing the dependent variable (commonly referred to as the responsibilities) of all observations. It is possible to restrict the model as not to include one or both of the Z and EV states if desirable. If the EV state is not included the model is equivalent of the ZINB model. The code for estimating the EVZINB model is written in R (**Rcore**). For prediction it is common to use the expected value of the dependent variable, given the covariates and the estimated parameters. However, the EV state lacks expected value if the shape parameter $\alpha_{EV,i} \leq 1$. Instead we use the harmonic mean for the EV state. The harmonic mean of a random variable X is defined as $E\left[X^{-1}\right]^{-1}$ and exists for any $\alpha_{EV,i} > 0$ for the Pareto distribution, although more conservative than $E[X]$. The predictions are used for investigating marginal effects. For details, see appendix A1 and A2.

In order to investigate whether a covariate has a significant contribution in any part of the model, a likelihood ratio (LR) test is developed. Define the vector of parameters to be $\boldsymbol{\theta}$ and the log likelihood to be $\ell(\boldsymbol{\theta})$ with the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ which maximizes $\ell(\boldsymbol{\theta})$. Define the parameter vector $\tilde{\boldsymbol{\theta}}$ which includes the restricted elements $\gamma_{Z,p} = \gamma_{EV,p} = \mu_{NB,p} = \alpha_{EV,p} = 0$ where for example $\gamma_{Z,p}$ represents the element of $\boldsymbol{\gamma}_Z$ corresponding to the p^{th} covariate in $\boldsymbol{\gamma}_Z$. Then the test statistic

$$\chi^2 = -2\left(\ell\left(\tilde{\boldsymbol{\theta}}\right) - \ell\left(\hat{\boldsymbol{\theta}}\right)\right)$$

asymptotically follows a χ^2 distribution with the degrees of freedom equal to the number of restrictions. In this case maximally 4 restrictions. It is possible to choose not to include any covariate in any part of the model.

In summary, the proposed method provides a tool for estimating a regression model with both excess zeroes and extreme values by maximum likelihood. In addition, bootstrap parameter distributions are provided.

4 Simulation Study

In order to investigate the performance of the proposed method, a simulation study is conducted. Data of three covariates were generated and the parameters $\boldsymbol{\gamma}_Z, \boldsymbol{\gamma}_{\boldsymbol{EV}}, \boldsymbol{\beta}_{NB}, \alpha_{NB}, \boldsymbol{\beta}_{EV}$ and c_{EV} are specified. 1000 replications of sample size $n = 1000$ are generated. The details of the simulation design is provided in the appendix A2.

4.1 Simulation Results

Table 2 shows the average parameter estimates using EVZINB, ZINB and NB where data is generated using three different data generating processes. From left to right data is generated using model parameters yielding a EV states of approximately 13.5%, 1.9% and 0.7%, respectively. While EVZINB provides average

[3] See appendix A1 for details on the Pareto distribution.

estimates close to the population values both estimates obtained using ZINB and NB are substantially biased. A less pronounced EV state leads to parameter estimates closer to true values using ZINB in general, as expected, although even the smallest EV state leads to substantial bias. For the case with no EV state, not included in the table, the EVZINB model is estimated restricting the EV state to be zero. Then the estimates are the same using ZINB and EVZINB models as expected.

Table 2. Population value (θ) and average estimates using EVZINB, ZINB and NB with three data-generating processes corresponding to $\gamma_{EV,0} = -1$ (yielding a EV state of approximately 13.5%), $\gamma_{EV,0} = -3$ (yielding a EV state of approximately 1.9%), and $\gamma_{EV,0} = -4$ (yielding a EV state of approximately 0.7%).

		13.5% EV			1.9% EV			0.7% EV		
	θ	EVZINB	ZINB	NB	EVZINB	ZINB	NB	EVZINB	ZINB	NB
$\beta_{NB,0}$	3	2.993	4.358	3.651	2.997	3.351	2.493	2.977	3.146	2.185
$\beta_{NB,1}$	0.2	0.198	0.273	0.484	0.200	0.272	0.454	0.202	0.172	0.363
$\beta_{NB,2}$	0.2	0.204	-0.078	-0.286	0.195	0.033	-0.154	0.201	0.171	-0.023
$\beta_{NB,3}$	0.2	0.203	0.269	0.471	0.205	0.272	0.455	0.200	0.168	0.355
α_{NB}	1	0.995	2.777	9.274	1.000	2.145	9.349	0.973	1.355	8.838

Table 6 in appendix A2 shows the standard standard deviation of parameter estimates using the three methods demonstrating that the EVZINB model provides the most efficient estimates in the investigated setting.

5 Empirical Study

To illustrate the empirical usefulness of the EVZINB regression model we replicate Moore's (2019) study on foreign fighters, social embeddedness, and violence against civilians. This study focuses on the effect of the presence and characteristics of *foreign fighters* on the annual observed number of fatalities from *one sided violence against civilians* (Eck and Hultman, 2007) by all rebel groups in 1989-2011, with a number of different model specifications where the main explanatory variable is re-specified and different control variables are used. Moore's main finding is that the presence of foreign fighters in a civil war increases the amount of violence directed against civilians, but that this effect varies depending on whether these fighters are coethnics of the rebel group, and depending on the distance the fighters travel.

We replicate Moore's negative binomial model with all control variables[4] (model 4a) and its zero-inflated counterpart from the online appendix (model 5a). We use bootstrapped estimates for all models as this allows for a comparison of the models on equal terms. In addition, using bootstrapped estimates of the coefficients allows us to compare the densities of the parameters and marginal effects without any distributional assumptions. A standard parametric regression table is presented in table 8 in appendix B1. Since each covariate may express its effect in four different parts of the EVZINB model, we have also developed and included a likelihood ratio test which tests the null hypothesis that there is a joint zero effect of the covariates on the dependent variable across all parts of the EVZINB model. The results of these likelihood ratio tests can be found in 9 in appendix B1.

Our aim with this replication is to show that re-estimating the model using our EVZINB regression allows for a more nuanced analysis of the effects of covariates on the outcome, and that EVZINB regression model outperforms the NB and ZINB regression models on a number of crucial metrics. The models we replicate contain 12 covariates for both the negative-binomial (count) estimation and for the estimation of zero-

[4] One of the covariates in the model, the lagged dependent variable was transformed from the count to the natural logarithm of the variable + 1 (as log(0) is undefined) in the analysis as this variable caused both models to be inestimable for a large proportion of the bootstrapped samples. This, in turn, suggests that the distributional characteristics of the variable makes it an unsuitable variable in the analysis

inflation, which means that the estimated number of parameters is 14 for the NB model[5] and 27 for the ZINB model[6].

For the EVZINB model we need to specify four sets of covariates. First we need to assign two sets of covariates to the zero-inflation and extreme value-inflation multinomial process which estimates the likelihood of observations belonging to the three states, zero, count, and extreme, of the model. We then assign a set of covariates to the Negative Binomial, count state, which just as a regular NB or ZINB model estimates the effect of covariates on the outcome, but weighted on the likelihood that the observations are from the count process. Lastly we assign one set of covariates to the extreme value state which are believed to influence the 'extremeness' of the values produced in the extreme value state.

For the count and zero-inflation estimation ($\boldsymbol{\gamma_Z}$ and $\boldsymbol{\gamma_{EV}}$) we have kept the same 12 covariates as in Moore's original NB and ZINB models so that we can make an immediate and fair comparison between the three models. For the estimation of extreme value inflation ($\boldsymbol{\gamma}_{EV}$), we removed one covariate, 'islamist conflict'. due to a lack of variation, so that extreme value inflation is estimated using 11 covariates. For the estimation of the Pareto shape parameter α_{EV}, governing the 'extremeness' of the extreme values, we chose to only include three covariates; the dummy variable on the presence of foreign fighters (*foreign_fighters*), the dummy variable for territorial conflict (*territorial*), and the natural logarithm of the lagged dependent variable (*log(fatalities_lag)*) as we believe that these covariates are the theoretically most relevant for the shape of the Pareto distribution.[7] The reason for limiting the number of covariates on the Pareto shape α_{EV} to three is that the amount of data in the extreme value state is relatively limited and an inclusion of many covariates risk leading to unstable estimation.

For a more detailed description of the variables and their summary statistics, see the online appendix for Moore's original article (Moore, 2019). In the analysis below, we have chosen to focus on the three covariates which are part of all three states (including both the extreme value inflation and Pareto shape in the EV state) in the EVZINB-model, in order to highlight the most substantive differences between the models. The full results and comparison between the models can be found in appendix B1.

5.1 Results

To analyze the results of the models we may either look at the bootstrapped parameter densities for the covariates, the equivalent of the results in a regression table, or look directly at the predicted values for the dependent variable for different values of the covariates in the model. Figure 1 below shows the mean predicted value and the mean predicted 90th percentile when setting the value of the *foreign fighters* covariate for all observations to 0 and 1 respectively, while keeping all other covariates at their observed values. As we expect the covariates to have different effects on the extreme and non-extreme values, showing both the mean predicted value and the mean of the 90th percentile allows for an interesting contrast between the effect of the covariates on the mean cases, and the effect on the most extreme cases. As an interpretation guideline, we would in practice expect 10% of observations to exceed the value in the mean 90th percentile plot, which means that we could interpret this plot as the effect of the covariates on the most extreme cases. When speaking about topics where extreme values are present, and thus where the EVZINB model would be a relevant modeling option, this is a highly relevant statistic as it shows the risk that the process enters into a risk-zone where very large counts become more likely.

[5] 12 covariates, 1 intercept, and the dispersion parameter

[6] 12 covariates and 1 intercept per state, and the dispersion parameter

[7] This brings the total number of estimated parameters for the EVZINB model to 44; 12 covariates each for the negative binomial and zero states, 11 covariates for extreme value inflation, 3 covariates for the Pareto shape α_{EV}, 4 intercepts, the dispersion parameter, and the cutoff-value, c_{EV}, for the Pareto distribution. In practice, we would suggest to make a theoretically informed choice of covariates for each state to reduce the total number of estimated parameters.

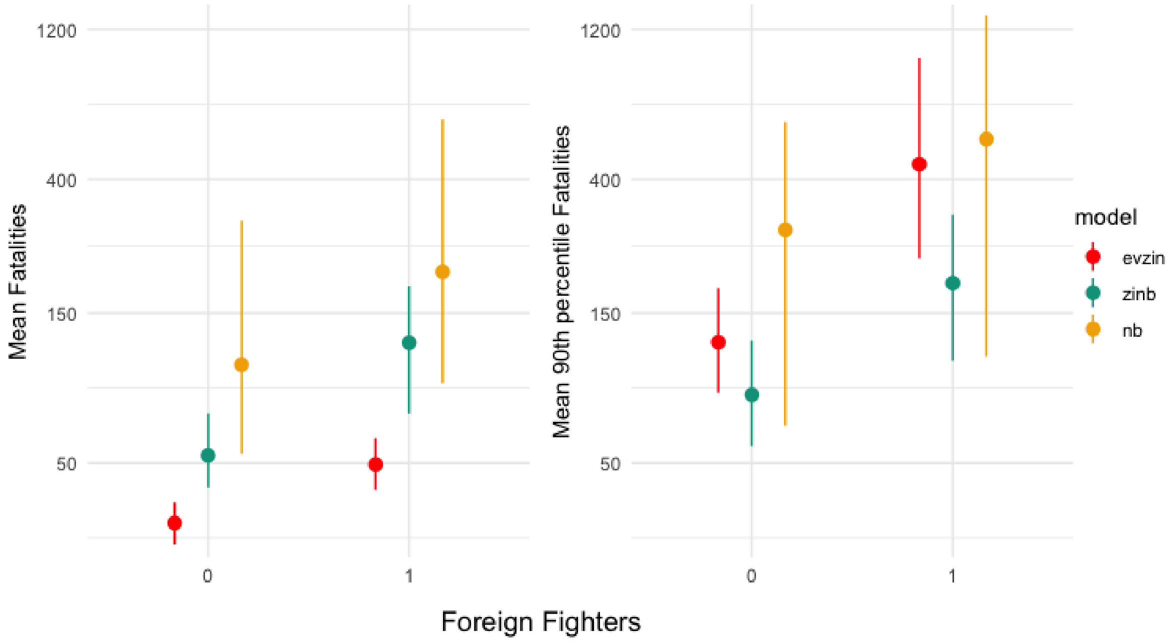

Figure 1. Mean predicted fatalities and mean 90th percentile for different values of *Foreign Fighters* across all bootstrapped samples with the remaining covariates set to their observed value, with 95% bootstrapped intervals.

The results in figure 1 highlight some important differences between the three models. First, we can see that all three models show an increase in the predicted mean number of OSV fatalities when foreign fighters are present in the rebel group. However, the form of this increase is very different. All three models agree that when foreign fighters are present both the predicted mean OSV count and the predicted 90th percentile of OSV counts increase substantially. Yet, in the NB model there is considerable uncertainty about both the mean and mean 90th percentile estimates and also a strong effect of foreign fighters on both. The ZINB model offers substantially more precise estimates for the mean OSV fatalities, but has a very similar effect on the mean of fatalities as on the mean 90th percentile of fatalities. The results from the EVZINB model, on the other hand, shows a more complex picture on the relationship between the presence of foreign fighters and the perpetration of OSV from rebel groups. For while the EVZINB model also shows an increase in the mean predicted OSV counts when foreign fighters are present, this increase is relatively modest, approximately from 30 to 50. Instead, the main effect of foreign fighters is seen in the predicted mean 90th percentile, where the EVZINB model shows a large increase, approximately from 120 to 450, with non-overlapping confidence intervals. This indicates that unlike the conclusions we can draw from the NB or ZINB models, that foreign fighters are simply associated with a higher expected count of OSV, the EVZINB model indicates that there is only a small effect of foreign fighters on the mean case of rebel OSV. Instead, the EVZINB model shows that the presence of foreign fighters primarily has an effect on the most extreme cases of rebel OSV. This means that the presence of foreign fighters can be seen as a risk-factor of rebel OSV spinning out of control, rather than a factor affecting the average cases.

A more disaggregated view of the same results can be seen in the plot of bootstrapped parameter densities across all three states of the EVZINB model, and the corresponding NB and ZINB parameter densities at their appropriate states, in figure 2 below. The means of these densities can roughly be interpreted as the beta coefficients in a regular regression framework, and an accompanying regression output table can be found in table 8 in the appendix. The patterns identified above regarding the *foreign fighters* covariates are confirmed by this plots and show that while there is very little effect of these covariates on $\boldsymbol{\beta}_{NB}$, where the densities are centered around zero, it can clearly be seen that $\boldsymbol{\gamma}_{EV}$ is almost exclusively positive and $\boldsymbol{\beta}_{EV}$

almost exclusively negative for foreign fighters, indicating that the presence of foreign fighters is associated with a higher likelihood of entering the extreme value domain ($\boldsymbol{\gamma}_{EV}$), and that when entering the extreme value domain the distributions tend to be 'more extreme' ($\boldsymbol{\beta}_{EV}$). These results are in line with the results from the likelihood ratio test of a joint zero effect of foreign fighters which was rejected at the 5% significance level. The results of the likelihood ratio tests for all parameters can be found in table 9 in the appendix.

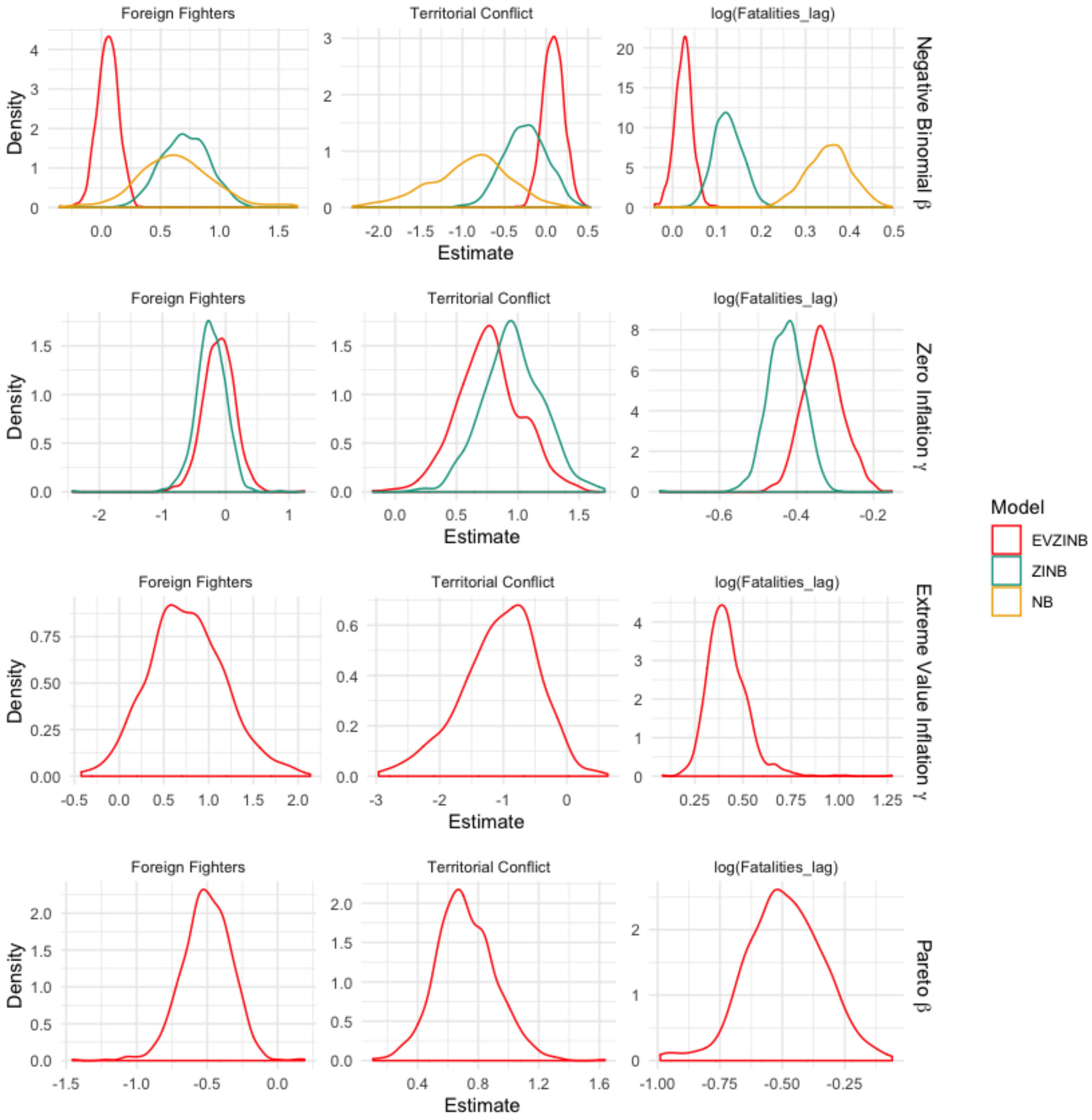

Figure 2. Density plots for the *foreign_fighters*, *territorial*, and *log(fatalities_lag)* covariates across all three states of the EVZINB model.

We also decided to look closer at one of the continuous covariates in the model, the log of the lagged dependent variable. Looking at figure 2 above, we can again see that in the NB and ZINB models $\boldsymbol{\beta}_{NB}$ for this covariate is almost exclusively positive, while in the EVZINB model it is centered much closer to zero with a substantial proportion below zero too. This would indicate a very small effect of the covariate on the counts in the negative binomial sub-process. However, looking at the other parameters of the model it is clear that these parameters for the log of lagged fatalities is almost exclusively negative for $\boldsymbol{\gamma}_{EV}$ and $\boldsymbol{\beta}_{EV}$ while exclusively positive for $\boldsymbol{\gamma}_{EV}$. Again, this is in line with the results from the likelihood ratio tests, which rejects a joint zero effect of *log(fatalities_lag)* in the EVZINB model. These results indicate that an increase in the log of lagged fatalities is associated with a lower likelihood of entering the zero state, a higher likelihood of entering the extreme value state, and a more extreme distribution for the extreme value state. These aspects can clearly be seen when investigating the predicted mean OSV counts, the mean 90th percentile OSV counts, and the component probabilities when setting the value of the log of lagged fatalities to different values, which can be seen in figure 3 below.

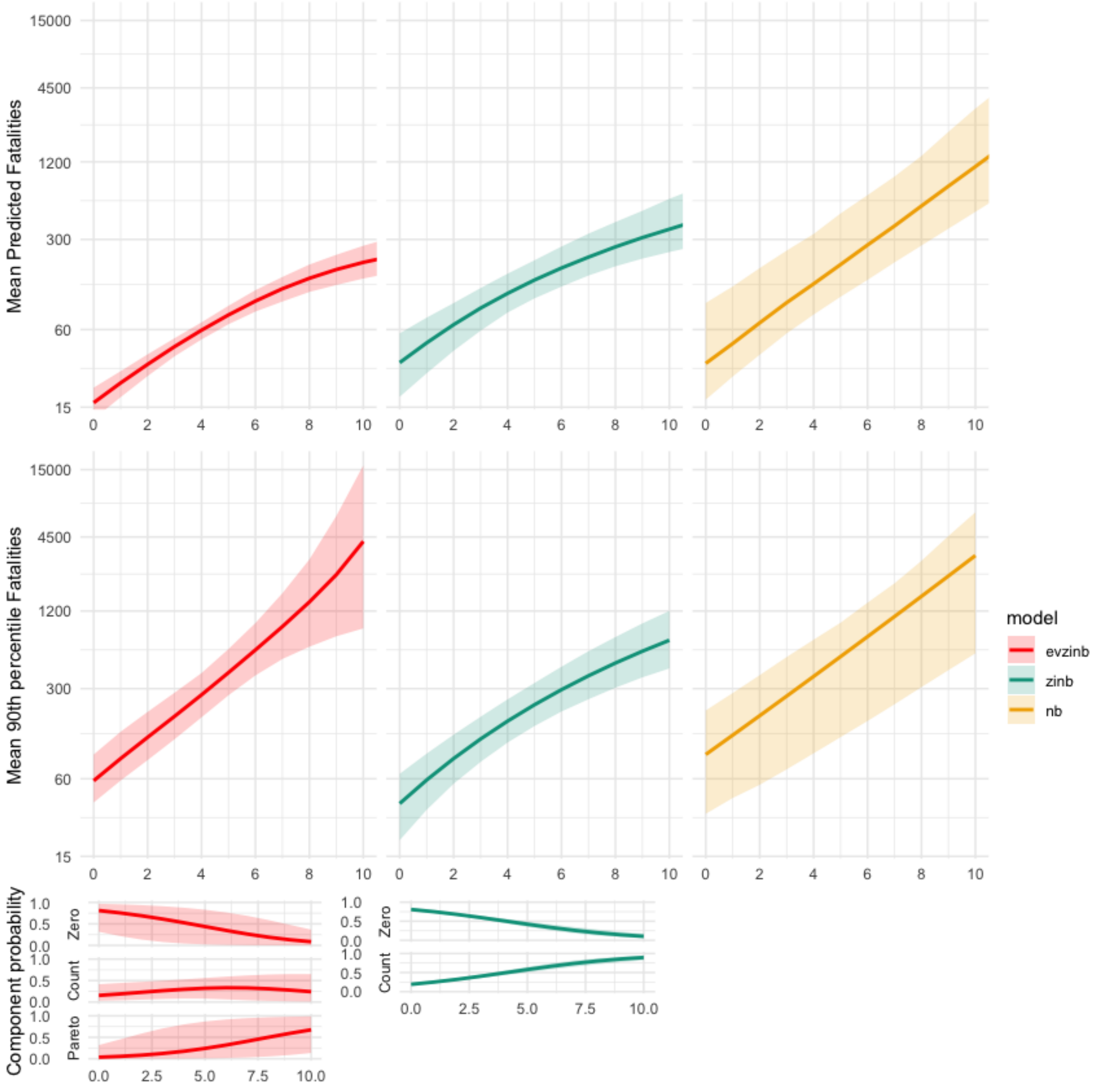

Figure 3. Mean predicted fatalities, mean 90th percentile, and state probabilities for different values of *log(fatalities_lag)* across all bootstrapped samples with the remaining covariates set to their observed value, with 95% bootstrapped intervals.

This figure confirms the patterns described above as it is clear that while all three models predict that the mean number of OSV counts increase with *log(fatalities_lag)* the changes are on quite different magnitudes. In the EVZINB model the mean predicted value from around 0 to just over 200 when *log(fatalities_lag)* increases from 0 to 10. Additionally, the EVZINB model also shows a non-monotone effect in the mean where the effect is tapering off as *log(fatalities_lag)* passes 6. The NB model, on the other hand, has an exponential effect of *log(fatalities_lag)* for the entire range seen as a straight line in the figure, while for the ZINB model the effect is also uniform over the entire range, albeit less than exponential. The ZINB and NB models also exhibit considerably more uncertainty regarding this mean estimate, especially the NB model which has a 95% bootstrapped confidence interval for the mean OSV count when *log(fatalities_lag)* is set to the maximum value ranging from just over 500 to over 4500.

However, while this covariate has a relatively limited effect on the predicted mean of fatalities in the EVZINB model we can again see that the effect is large on the mean predicted 90th percentile where there is an increase from around 50 to over 4000 as the covariate increases from 0 to 10. The shape of this increase is also interesting as it in the rightmost part of the plot increases at an above exponential pace, seen when the line inches upward compared to a straight line. This confirms that in the EVZINB model, the effect of *log(fatalities_lag)* is primarily an effect on the most extreme values. This is also evident when looking at the state probabilities which clearly show that as *log(fatalities_lag)* increases, the likelihood of entering into the extreme value process increases too. These are the exact same patterns discussed in the results section above which indicated that *log(fatalities_lag)* increased the likelihood of entering into the extreme value state, and that increased *log(fatalities_lag)* also affected the Pareto shape to be more extreme. Comparing the results from the EVZINB model with the ZINB and NB models show some interesting variation. While the ZINB model shows a relatively small effect of *log(fatalities_lag)* on the 90th percentile, increasing from around 40 to just over 200 when *log(fatalities_lag)* increases from 0 to 10 the NB model shows a pattern similar to the EVZINB model, but without the inferential nuances which the EVZINB model brings out.

5.2 Model performance

While the section above highlights that the EVZINB model allows for a more nuanced analysis of effects than the NB and ZINB models, it is also important to test the performance of the models on a range of different metrics in order to determine the usefulness of the EVZINB model.

Model fit

To compare the fit of the NB, ZINB and EVZINB models we use a bootstrapped likelihood ratio test using Aikaike Information Criteria (AIC) or Bayesian Information Criteria (BIC) correction (Konishi and Kitagawa, 2008). This is similar to a bootstrapped extension of Vuong's (1989) test for non-nested models. The results, which can be seen in figure 5 and table 10 in appendix B2, show that with AIC correction the EVZINB model outperforms the NB model in 100% and the ZINB model in 99.7% of the bootstrapped samples. Using a BIC correction, the EVZINB model still outperforms the NB model in 100% of the bootstrapped cases, and outperforms the ZINB model in 75% of of the bootstrapped cases. These results indicates that the EVZINB model fits the data substantially better than the competing models.

Another way of assessing how well the model fits the data is to compare the conditional densities produced by different models with the observed empirical distribution of values. Figure **??** in the appendix shows the conditional densities of the three models and of the empirical data. This figure clearly show that the conditional density produced by EVZINB model is substantially closer to the empirical density. .

Predictive performance

Apart from model fit, we also test the predictive performance of the ZINB and EVZINB models using the out of sample *root mean squared log error* (RMSLE) metric from 100 15-fold cross validations.[8] Out of the 1500 cross-validation samples there are 412 inadmissible samples using NB and no inadmissible samples using ZINB or EVZINB. The RMSLE was compared using the NB, ZINB and EVZINB models for the 1088 samples that converged for all models. EVZINB outperforms NB in 100% of the samples and ZINB in 96.1% of the samples. This clearly shows that the EVZINB model outperforms the ZINB model both with regards to its model fit, and to its predictive performance. These results indicate that the introduction of the EVZINB model is indeed both empirically and theoretically useful.

[8] The models are evaluated against the RMSLE rather than the more commonly used *root mean squared error* (RMSE) since the RMSLE is a measure of the *relative* error of the model. In data with extreme observations the RMSE can be highly sensitive to a few extreme observations.

6 Discussion

Section 5 of this paper has shown that the EVZINB model can be empirically and theoretically useful. By employing the EVZINB model instead of the more conventional NB or ZINB models we were able to disentangle how the effects of certain covariates affect not only the mean cases but also how the effects of these covariates differ between the effect on the mean and on the extreme cases.

We were able to show that the main effect of the presence of foreign fighters on rebel group use of OSV was not on the average cases, where the effect was rather modest, but instead on the most extreme cases as the presence of foreign fighters both increased the likelihood of experiencing extreme values and tended to make the extreme values more extreme. While this is in agreement with Moore's original conclusion that foreign fighters is an important factor for explaining OSV fatalities from rebel groups, the more nuanced inferences from the EVZINB model allows us to pinpoint that foreign fighters can be seen as a risk factor which may case a rebel group to enter into the extreme value domain where the risk of observing a very high number of OSV fatalities is high. This may prove useful information to policy makers who may ponder the risks and benefits of different course of action. Additionally, researchers can test more fine-grained hypotheses not only about which covariates affect which outcomes, but also test hypotheses relating to the extreme and less extreme cases. There may, for instance, be covariates which both decrease the likelihood of observing a zero, but also decreases the likelihood of entering the extreme value domain. With conventional models such as the NB and ZINB models, hypotheses relating to these covariates would be difficult to test as the mean effects may be zero, even if the effects are pronounced in both directions. In addition to providing new and useful insights in the empirical case, we also showed that the EVZINB model outperformed both the NB and the ZINB model on a number of different metrics when extreme values are present in the analysis.

In the analysis above we have chosen to theoretically analyze all states of the model, however this is not necessary for the EVZINB model to be useful. Rather, the EVZINB model can be used in a number of other settings as well. In the simplest setting, the introduction of the EV state of the EVZINB model can be seen as a filter for the extreme values included in this analysis. Viewed through this perspectives, the EVZINB model would simply allow for an unbiased and more stable, or efficient, estimation of the *count* process. This is the equivalent of the filtering function of any zero-inflated model, where no covariates or analysis is focused on the zero state, but rather the zero-state is included solely to filter out the effect of the 'excess zeroes'. By only estimating an intercept for the EV state of the EVZINB model, the EVZINB model would then filter out the 'excess' effect of extreme observations in the data, allowing the researcher to focus their analysis on the count process.

Another way of viewing the EV state of the EVZINB model is to allow covariates to affect the Pareto parameter α_{EV} of the process, i.e. to model the effect of different covariates on the 'extremeness' of the extreme values. For this approach, we would theoretically motivate a number of covariates which would be expected to have an effect on the analyzed outcome in these *extreme* cases, which would then be evaluated much in the same way as regular regression coefficients. This approach would allow researchers to investigate the direction and size of the effect of covariates on the outcome variable for the different latent states. This may be of particular interest when some covariates may be assumed to have heterogeneous or divergent effects depending on the latent state. The marginal effects of certain covariates over the different latent states may also be investigated, to provide a more nuanced, and less linear, picture of how the dependent and independent variables relate.

Additionally, regression parameters may also be estimated for the Pareto threshold, $\hat{c}_{EV}$, i.e. the point after which observations may enter into the latent Pareto process. While perhaps not immediately evident, estimating regression parameters specifically for $\hat{c}_{EV}$ may create new avenues for research, and open up new types of research questions as $\hat{c}_{EV}$ is a measure for *when* a process have a chance of progressing from a well-behaved count process to an extreme value process. This means that by investigating covariates effect on $\hat{c}_{EV}$ we can ask questions such as: '*what factors affect the threshold for when low intensity armed conflicts may progress into large scale armed conflicts?*'.

Similarly, investigating the covariates for the multinomial process which differentiates between the zero, count, and extreme value latent states, will also open up research question on what factors affect the probability of a process graduating from one latent state to another, for instance from the well-behaved count process to the unpredictable extreme value process. To aid in analysis of these questions, the predicted ex-ante state probabilities, i.e. the probability of the different latent states given the covariates of an observation, and the ex-post probabilities, i.e. the probability of the different latent states given the covariates *and* the observed dependent variable of an observation, may be of great use as they provide an opportunity to assess the status of the process both prior to and after observing the dependent variable. This may be especially helpful in guiding forecasting efforts, and may be useful when designing policy recommendation, as it provides both guidance and evaluation metrics. For instance, the state probabilities may be used to evaluate the risk of a low-intensity armed conflict escalating into a high-intensity armed conflict, while the ex-post probabilities of the same observation may be used to assess the likelihood that the armed conflict, given a certain number of fatalities, is in the extreme value domain.

While the EVZINB model has a lot of qualities which makes it a useful model in a range of different empirical applications, there are also some limitations. Not least among these are the fact that there usually are relatively few observations with extreme values in a dataset. This lack of information may make it difficult to reach strong conclusions about certain covariates effect for the different parts of the Pareto process. Additionally, using a large number of covariates on each of the estimable parts will cause the total number of parameters in the model to be very high and may cause overfitting and unstable estimation. Our suggestion is to be carefully parsimonious and only use theoretically motivated covariates in each state. In addition, we suggest the use of bootstrapped parameter corrected (AIC or BIC) likelihood ratio tests to test different specifications of the EVZINB model to alternatives such as the NB or ZINB models. It is also worth noting that while we have introduced the full EVZINB model in this paper, it is also possible to use this model in data which are *extreme value inflated* without being zero inflated.

7 Conclusion

In this paper we have introduced the extreme value and zero inflated regression model for count data which contains both an inflated number of zeroes and extreme values. The extreme value and zero inflated regression model can be thought of as a latent states regression model, where we can estimate both which covariates affect the likelihood of different states of the process, and how these covariates affect the behavior of the process given its state. This allows us to properly model data which can be thought of as having been generated from different data generating processes.

We have shown that this model is both empirically and theoretically motivated, and that the model can retrieve correct parameter estimates from simulated data. We have also shown the empirical usefulness of the model through a replication of a recently published study on the use of one sided violence against civilians by rebel groups. In the replication study, the extreme value and zero inflated regression model allowed the researcher to draw inferences which were not available when using a negative binomial or zero-inflated negative binomial model. Additionally, the EVZINB model outperformed both the NB and ZINB models with regards to efficiency of the estimated parameters, the AIC and BIC corrected likelihood ratios, and the predictive performance of the model.

We have also presented a number of different empirical lenses through which the extreme value and zero inflated regression model can be viewed, and how this model can allow researchers to ask novel questions about the nature of their data, and to ask questions previously not possible to answer. The extreme value and zero inflated regression model can also easily be extended to a non-zero inflated version, for count data which do not suffer from zero inflation, but still contains extreme values. With future development a unified framework for analysis of the effect of covariates across states of the model could be developed, allowing for a more specific analysis of the marginal effects of certain covariates in different conditions.

The extreme value and zero inflated negative binomial model and tools related to analyzing this model will, in the future, be available in the `R` package `evi`.

Funding

This work was supported by the European Research Council [grant number H2020-ERC-2015-AdG 694640 to D.R.], and the Swedish Research Council [grant number 2017-01175 to V.J.]. For more information on the project see http://www.pcr.uu.se/research/views/.

Acknowledgements

Equal authorship is implied. We would like to thank Lovisa Vildö, Håvard Hegre, Espen Geelmuyden Rød, Shaobo Jin, Fan Yang Wallentin, and all the members of the Monday Group on Wedensdays for thoughtful comments and invaluable support throughout the writing process.

References

Bagozzi, Benjamin E. (Sept. 2015). "Forecasting Civil Conflict with Zero-Inflated Count Models". en. In: *Civil Wars*. Publisher: Routledge.

Cancho, Ramon Ferrer i and Ricard V Solé (2003). "Least effort and the origins of scaling in human language". In: *Proceedings of the National Academy of Sciences* 100.3, pp. 788–791.

Cirillo, Pasquale and Nassim Nicholas Taleb (2016). "On the statistical properties and tail risk of violent conflicts". In: *Physica A: Statistical Mechanics and its Applications* 452, pp. 29–45.

Clauset, Aaron (Sept. 2017). *The Enduring Threat of a Large Interstate War*. Tech. rep. One Earth Future Foundation.

Clauset, Aaron, Maxwell Young, and Kristian Skrede Gleditsch (Feb. 2007). "On the Frequency of Severe Terrorist Events". en. In: *Journal of Conflict Resolution* 51.1, pp. 58–87.

Cunen, Céline, Nils Lid Hjort, and Håvard Mokleiv Nygård (2020). "Statistical sightings of better angels: Analysing the distribution of battle-deaths in interstate conflict over time". In: *Journal of peace research* 57.2, pp. 221–234.

Dean, Charmaine and Jerald Franklin Lawless (1989). "Tests for detecting overdispersion in Poisson regression models". In: *Journal of the American Statistical Association* 84.406, pp. 467–472.

Dempster, Arthur P, Nan M Laird, and Donald B Rubin (1977). "Maximum likelihood from incomplete data via the EM algorithm". In: *Journal of the royal statistical society. Series B (methodological)* 39, pp. 1–38.

Disha, Ilir (2019). "Different Paths: The Role of Immigrant Assimilation on Neighborhood Crime*". en. In: *Social Science Quarterly* 100.4. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ssqu.12618, pp. 1129–1153.

Dixon, W. J. and K. K. Yuen (June 1974). "Trimming and winsorization: A review". en. In: *Statistische Hefte* 15.2, pp. 157–170.

Eck, K. and L. Hultman (Mar. 2007). "One-Sided Violence Against Civilians in War: Insights from New Fatality Data". en. In: *Journal of Peace Research* 44.2, pp. 233–246.

Frome, Edward L, Michael H Kutner, and John J Beauchamp (1973). "Regression analysis of Poisson-distributed data". In: *Journal of the American Statistical Association* 68.344, pp. 935–940.

Greene, William H (1994). "Accounting for excess zeros and sample selection in Poisson and negative binomial regression models". In:

Gutenberg, Beno and Charles F Richter (1944). "Frequency of earthquakes in California". In: *Bulletin of the Seismological Society of America* 34.4, pp. 185–188.

Hegre, Håvard et al. (2019). "ViEWS: a political violence early-warning system". In: *Journal of peace research* 56.2, pp. 155–174.

Hultman, Lisa (Jan. 2013). "UN peace operations and protection of civilians Cheap talk or norm implementation?" en. In: *Journal of Peace Research* 50.1, pp. 59–73.

Konishi, Sadanori and G. Kitagawa (2008). *Information criteria and statistical modeling*. en. Springer series in statistics. New York: Springer.

Lacina, Bethany and Nils Petter Gleditsch (June 2005). "Monitoring Trends in Global Combat: A New Dataset of Battle Deaths". en. In: *European Journal of Population / Revue européenne de Démographie* 21.2, pp. 145–166.

Lange, Kenneth (1995). "A gradient algorithm locally equivalent to the EM algorithm". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 57.2, pp. 425–437.

Lawless, Jerald F (1987). "Negative binomial and mixed Poisson regression". In: *Canadian Journal of Statistics* 15.3, pp. 209–225.

Litvinenko, Yuri E (1996). "A new model for the distribution of flare energies". In: *Solar Physics* 167.1-2, pp. 321–331.

Liu, Chuanhai and Donald B Rubin (1994). "The ECME algorithm: a simple extension of EM and ECM with faster monotone convergence". In: *Biometrika* 81.4, pp. 633–648.

Mandelbrot, Benoit B (1983). *The fractal geometry of nature*. Vol. 173. WH freeman New York.

Minami, Mihoko et al. (2007). “Modeling shark bycatch: the zero-inflated negative binomial regression model with smoothing”. In: *Fisheries Research* 84.2, pp. 210–221.

Moore, Pauline (Mar. 2019). “When do ties bind? Foreign fighters, social embeddedness, and violence against civilians”. In: *Journal of Peace Research* 56.2. Publisher: SAGE Publications Ltd, pp. 279–294.

Pettersson, Therese and Magnus Öberg (2020). “Organized violence, 1989–2019”. In: *Journal of peace research* 57.4, pp. 597–613.

Vuong, Quang H. (1989). “Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses”. In: *Econometrica* 57.2. Publisher: [Wiley, Econometric Society], pp. 307–333.

Weidmann, Nils B. and Espen Geelmuyden Rød (2019). *The Internet and Political Protest in Autocracies*. en. Google-Books-ID: 0iaeDwAAQBAJ. Oxford University Press.

Wu, CF Jeff (1983). “On the convergence properties of the EM algorithm”. In: *The Annals of statistics*, pp. 95–103.

Appendix A1: Statistical methodology

Likelihood

It is assumed that each observation i has been generated from one of the three data generating processes Z, NB and EV, representing the three latent states, with respective prior probabilities $\pi_{Z,i}$, $\pi_{NB,i}$, and $\pi_{EV,i}$, with the restriction $\pi_{Z,i} + \pi_{NB,i} + \pi_{EV,i} = 1$. The corresponding probability mass functions are denoted $f_Z(\cdot)$, $f_{NB}(\cdot)$ and $f_{EV}(\cdot)$, respectively. The probability mass function of the i^{th} observation is

$$f(y_i) = \pi_{Z,i} f_Z(y_i) + \pi_{NB,i} f_{NB}\left(y_i; \mu_{NB,i}, \alpha_{NB}\right) + \pi_{EV,i} f_{EV}\left(y_i; \alpha_{EV,i}, c_{EV}\right), \tag{4}$$

where $f_Z(y_i) = \delta_{y_i,0}$, defined by

$$\delta_{y_i,0} = \begin{cases} 1, & y_i = 0 \\ 0 & y_i \neq 0 \end{cases}.$$

Further,

$$f_{NB}\left(y_i; \mu_{NB,i}, \alpha_{NB}\right) = \frac{\Gamma\left(y_i + \alpha_{NB}^{-1}\right)}{\Gamma(y_i + 1)\Gamma\left(\alpha_{NB}^{-1}\right)} \left(\frac{1}{1 + \mu_{NB,i}\alpha_{NB}}\right)^{\alpha_{NB}^{-1}} \left(1 - \frac{1}{1 + \mu_{NB,i}\alpha_{NB}}\right)^{y_i}, \tag{5}$$

where $\Gamma(\cdot)$ is the gamma function, $\mu_{NB,i}$ is the expected value of observation i conditioned on the NB latent state and $\alpha_{NB} > 0$ is the corresponding dispersion parameter. A random variable with the probability mass function as in Equation (5) has variance

$$V(Y_i) = \mu_{NB,i}\left(1 + \alpha_{NB}\mu_{NB,i}\right).$$

In the limit as α_{NB} approaches 0, the variance approaches $\mu_{NB,i}$ and the distribution approaches the Poisson distribution. The EV state is modeled as a discretized version of the Pareto distribution. The probability density function of the Pareto distribution is

$$f(y; \alpha, c) = \mathbf{1}_{y \geq c}\left(\frac{\alpha c^{\alpha}}{y^{\alpha+1}}\right), \;\; y > c, \;\; \alpha > 0, \;\; c > 0, \tag{6}$$

where c is a scalar, representing the lower bound for the domain of the random variable and α is the shape parameter, controlling the shape of the tail of the distribution. $\mathbf{1}_{y \geq c}$ is the indicator function, defined as

$$\mathbf{1}_{y \geq c} = \begin{cases} 1, & y \geq c \\ 0 & y < c \end{cases}.$$

The distribution according to equation (6) is that of a continuous random variable. In the context of count data, a discrete version of the Pareto distribution is appropriate. The corresponding probability mass function of observation i is then

$$f_{EV}\left(y_i : \alpha_{EV,i}, c_{EV}\right) = \mathbf{1}_{y_i \geq c_{EV}} \left[\left(\frac{c_{EV}}{y_i}\right)^{\alpha_{EV,i}} - \left(\frac{c_{EV}}{y_i+1}\right)^{\alpha_{EV,i}} \right], \tag{7}$$

where c_{EV} is the lower bound for Y_i conditioned on the EV latent state, and $\alpha_{EV,i}$ is the corresponding shape parameter. Write the probability mass function in equation (4) as $f\left(y_i\right) = f\left(y_i; \boldsymbol{\theta}\right)$, where

$$\boldsymbol{\theta}^T = \left(\boldsymbol{\gamma}_Z^T, \boldsymbol{\gamma}_{EV}^T, \boldsymbol{\beta}_{NB}^T, \boldsymbol{\beta}_{EV}^T, \alpha_{NB}, c_{EV}\right).$$

The log likelihood is then

$$\mathcal{L}\left(\boldsymbol{\theta}\right) = \sum_{i=1}^{n} \log f\left(y_i; \boldsymbol{\theta}\right), \tag{8}$$

where n is the number of observations. The vector of parameters $\hat{\boldsymbol{\theta}}$ that maximizes equation (8) is the maximum likelihood estimator of $\boldsymbol{\theta}$, i.e.,

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\arg\max} \mathcal{L}\left(\boldsymbol{\theta}\right), \tag{9}$$

where $\boldsymbol{\Theta}$ is the domain of $\boldsymbol{\theta}$.

Estimation

An expectation-maximization (EM) approach of Dempster, Laird, and Rubin, 1977 is used to estimate $\boldsymbol{\theta}$ in equation (9). Start by introducing another random variable, $\boldsymbol{W}_i$ of the i^{th} observation. $\boldsymbol{W}_i$ has possible outcomes:

- If $\boldsymbol{W}_i^T = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix}$, Y_i is generated from the Z latent state.
- If $\boldsymbol{W}_i^T = \begin{pmatrix} 0 & 1 & 0 \end{pmatrix}$, Y_i is generated from the NB latent state.
- If $\boldsymbol{W}_i^T = \begin{pmatrix} 0 & 0 & 1 \end{pmatrix}$, Y_i is generated from the EV latent state.

Let $\boldsymbol{w}_i$ be a realization of $\boldsymbol{W}_i$, and $w_{i,j}$ be the j^{th} ($j = \{1,2,3\}$) state of $\boldsymbol{w}_i$. Then define the joint probability density of $\boldsymbol{W}_1, \boldsymbol{W}_2, ..., \boldsymbol{W}_n$ and $Y_1, Y_2, ..., Y_n$ as $e^{\mathcal{L}_c(\boldsymbol{\theta})}$ or the complete likelihood. Then the complete log likelihood is

$$\mathcal{L}_c\left(\boldsymbol{\theta}\right) = \sum_{i=1}^{n} w_{i,1} \log\left\{\pi_{Z,i} f_Z\left(y_i; \boldsymbol{\theta}\right)\right\} + w_{i,2} \log\left\{\pi_{NB,i} f_{NB}\left(y_i; \boldsymbol{\theta}\right)\right\} + w_{i,3} \log\left\{\pi_{EV,i} f_{EV}\left(y_i; \boldsymbol{\theta}\right)\right\}. \tag{10}$$

Note that the ex-post probabilities of $\boldsymbol{W}_i$, after observing Y_i, are

$$\begin{aligned}
\Pr\left(W_{i,1} = 1 | y_i\right) &= \frac{\pi_{Z,i} f_Z\left(y_i; \boldsymbol{\theta}\right)}{\pi_{Z,i} f_Z\left(y_i; \boldsymbol{\theta}\right) + \pi_{NB,i} f_{NB}\left(y_i; \boldsymbol{\theta}\right) + \pi_{EV,i} f_{EV}\left(y_i; \boldsymbol{\theta}\right)} \\
\Pr\left(W_{i,2} = 1 | y_i\right) &= \frac{\pi_{NB,i} f_{NB}\left(y_i; \boldsymbol{\theta}\right)}{\pi_{Z,i} f_Z\left(y_i; \boldsymbol{\theta}\right) + \pi_{NB,i} f_{NB}\left(y_i; \boldsymbol{\theta}\right) + \pi_{EV,i} f_{EV}\left(y_i; \boldsymbol{\theta}\right)} \\
\Pr\left(W_{i,3} = 1 | y_i\right) &= \frac{\pi_{EV,i} f_{EV}\left(y_i; \boldsymbol{\theta}\right)}{\pi_{Z,i} f_Z\left(y_i; \boldsymbol{\theta}\right) + \pi_{NB,i} f_{NB}\left(y_i; \boldsymbol{\theta}\right) + \pi_{EV,i} f_{EV}\left(y_i; \boldsymbol{\theta}\right)}
\end{aligned} \tag{11}$$

Let $f\left(\boldsymbol{w}_i | y_i; \boldsymbol{\theta}^{(0)}\right)$ represent the probability mass function of $\boldsymbol{w}_i$ according to equation (11). Define the expected value of the complete log likelihood function, with the distribution of $\boldsymbol{W}_i$ conditioned on y_i for $i = \{1, 2, ..., n\}$ following the distribution according to $f\left(\boldsymbol{w}_i | y_i; \boldsymbol{\theta}^{(0)}\right)$, as

$$Q\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(0)}\right) = \mathbb{E}_{\boldsymbol{W}; \boldsymbol{\theta}^{(0)}}\left[\mathcal{L}_c\left(\boldsymbol{\theta}\right)\right],$$

where $\boldsymbol{W} = \{\boldsymbol{W}_1, \boldsymbol{W}_2, ..., \boldsymbol{W}_n\}$ and $\mathbb{E}_{\boldsymbol{W};\boldsymbol{\theta}^{(0)}}[\cdot]$ takes the expected value of the random variable it acts on, assuming the distribution of $\boldsymbol{W}$ is given by the parameter vector $\boldsymbol{\theta}^{(0)}$. This is the E-step in an EM algorithm. In the M-step, select the parameter vector $\boldsymbol{\theta}^{(1)}$ for which

$$\boldsymbol{\theta}^{(1)} = \underset{\boldsymbol{\theta}\in\Theta}{\arg\max} Q\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(0)}\right). \tag{12}$$

Partition $\boldsymbol{\theta}^T = \left(\widetilde{\boldsymbol{\theta}}^T, c_{EV}\right)$, where $\widetilde{\boldsymbol{\theta}}^T = \left(\boldsymbol{\gamma}_Z^T, \boldsymbol{\gamma}_{EV}^T, \boldsymbol{\beta}_{NB}^T, \boldsymbol{\beta}_{EV}^T, \alpha_{NB}\right)$. No closed-form solution exists for equation (12) with respect to either $\widetilde{\boldsymbol{\theta}}$ nor c_{EV}. Hence, a generalized EM (GEM) algorithm, based on the Newton-Raphson step, is proposed to update $\widetilde{\boldsymbol{\theta}}$. See for example Wu, 1983; Lange, 1995. However, $Q\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(0)}\right)$ is not continuous with respect to c_{EV}. Neither is it once, nor twice differentiable. Thus, the GEM algorithm is inapplicable with regards to c_{EV}. On the other hand, since $\mathcal{L}(\boldsymbol{\theta})$ is a monotone function of c_{EV} between unique values of y_i, its maximum (if it exists) is attained at at least one of the unique values. We therefore propose to, after updating $\widetilde{\boldsymbol{\theta}}$, update c_{EV} by choosing $c_{EV}^{(k+1)}$ that maximizes $\mathcal{L}\left(\widetilde{\boldsymbol{\theta}}^{(k+1)}, c_{EV}\right)$. This can be viewed as a version of the expectation conditional maximization either (ECME) algorithm of Liu and Rubin, 1994. Due to the discontinuity of $\mathcal{L}(\boldsymbol{\theta})$ with respect to c_{EV} bootstrap standard errors are used for inference in this work. The proposed procedure is summarized in the pseudocode below at iteration k. m is the number of Newton-Raphson steps in the GEM algorithm, η_l and η_u are lower and upper limits of the multiplicative constant η of the Newton-Raphson step. C is the set of all unique values of observed y_i where $i \in \{1, 2, ..., n\}$.

```
while convergence criterion not met do
    $\tilde{\boldsymbol{\theta}}_1^{(k)} \leftarrow \tilde{\boldsymbol{\theta}}^{(k)}$
    for $l$ in $1 : m$ do
        $\widetilde{\boldsymbol{\theta}}_{l+1}^{(k)} \leftarrow \tilde{\boldsymbol{\theta}}_l^{(k)} - \left[\frac{\partial^2}{\partial\widetilde{\boldsymbol{\theta}}\partial\widetilde{\boldsymbol{\theta}}^T} Q\left(\widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\theta}}_l^{(k)}, c_{EV}^{(k)}\right)\right]^{-1}\Bigg|_{\widetilde{\boldsymbol{\theta}}=\widetilde{\boldsymbol{\theta}}_l^{(k)}} \left[\frac{\partial}{\partial\widetilde{\boldsymbol{\theta}}} Q\left(\widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\theta}}_l^{(k)}, c_{EV}^{(k)}\right)\right]\Bigg|_{\widetilde{\boldsymbol{\theta}}=\widetilde{\boldsymbol{\theta}}_l^{(k)}}$
    end
    $\widetilde{\boldsymbol{\theta}}^{*(k+1)} \leftarrow \widetilde{\boldsymbol{\theta}}_{l+1}^{*(k)}$
    if $\mathcal{L}\left(\widetilde{\boldsymbol{\theta}}^{*(k+1)}, c_{EV}^{(k)}\right) < \mathcal{L}\left(\tilde{\boldsymbol{\theta}}^{(k)}, c_{EV}^{(k)}\right)$ then
        $\hat{\eta} \leftarrow \underset{\eta\in[\eta_l,\eta_u]}{\arg\max} \mathcal{L}\left(\widetilde{\boldsymbol{\theta}}^{*(k)} + \eta\left\{\widetilde{\boldsymbol{\theta}}^{*(k+1)} - \widetilde{\boldsymbol{\theta}}^{(k)}\right\}, c_{EV}^{(k)}\right)$
        $\widetilde{\boldsymbol{\theta}}^{(k+1)} \leftarrow \widetilde{\boldsymbol{\theta}}^{(k)} + \hat{\eta}\left\{\widetilde{\boldsymbol{\theta}}^{*(k+1)} - \widetilde{\boldsymbol{\theta}}^{(k)}\right\}$
    else
        $\widetilde{\boldsymbol{\theta}}^{(k+1)} \leftarrow \widetilde{\boldsymbol{\theta}}^{*(k+1)}$
    end
    $c_{EV}^{(k+1)} \leftarrow \underset{c_{EV}\in C}{\arg\max} \mathcal{L}\left(\widetilde{\boldsymbol{\theta}}^{(k+1)}, c_{EV}\right)$
end
```

Appendix A2: Simulations

Performance metrics in simulations

The bias of the p^{th} parameter is

$$\text{Bias}_p = \frac{1}{R}\sum_{r=1}^{R} \hat{\theta}_{p,r} - \theta_p \tag{13}$$

where $\hat{\theta}_{p,r}$ is the estimate of parameter θ_p in the r^{th} simulated sample, and R is the number of simulated samples. SD of the p^{th} parameter is

$$\mathrm{SD}_p = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\hat{\theta}_{p,r} - \bar{\theta}_p\right)^2} \tag{14}$$

where $\bar{\theta}_p$ is the average estimate of the parameter θ_p, and RMSE of the p^{th} parameter is

$$\mathrm{RMSE}_p = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\hat{\theta}_{p,r} - \theta_p\right)^2}.$$

Prediction

For prediction of the dependent variable Y it is most common to use the expected value. In the proposed three-state setting this gives

$$\widehat{Y} = E\left[Y\right] = \pi_{NB}\mu_{NB} + \pi_{EV}\mu_{EV}, \tag{15}$$

where π_{NB} and π_{EV} are the probabilities of the NB and EV states, respectively and μ_{NB} and μ_{EV} are the means conditioned on the respective DGPs. For a Pareto distribution with shape parameter α_{EV} and lower limit c_{EV}, the mean is undefined for $\alpha_{EV} < 1$. Thus, we propose to replace the mean by the harmonic mean, which is

$$\widetilde{\mu}_{EV} = \frac{1}{E\left[\frac{1}{Y}\right]} = \frac{c_{EV}\left(\alpha_{EV}+1\right)}{\alpha_{EV}}. \tag{16}$$

Equations (2), (3), (16), and (16) yield the suggested prediction

$$\widehat{Y} = \frac{\exp\left\{\boldsymbol{\beta}_{NB}\boldsymbol{x}\right\} + c_{EV}\exp\left\{\boldsymbol{\gamma}_{EV}^T\boldsymbol{x} - \boldsymbol{\beta}_{EV}^T\boldsymbol{x}\right\} + c_{EV}\exp\left\{\boldsymbol{\gamma}_{EV}^T\boldsymbol{x}\right\}}{1 + \exp\left\{\boldsymbol{\gamma}_Z^T\boldsymbol{x}\right\} + \exp\left\{\boldsymbol{\gamma}_{EV}^T\boldsymbol{x}\right\}}. \tag{17}$$

Simulation design

First, three covariates $\boldsymbol{x}_i$ for each observation $i = \{1, 2, ..., n\}$ are generated from a multivariate normal distribution with zero mean vector and covariance matrix

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1.0 & & \\ 0.4 & 1.0 & \\ 0.3 & 0.2 & 1.0 \end{pmatrix}.$$

Ones are added in the first element of every $\boldsymbol{x}_i$ corresponding to an intercept. Secondly, the latent state is selected among Z, NB and EV based on the probabilities $\pi_{Z,i}$ and $\pi_{EV,i}$ in equation (2). Given the latent state, the dependent variable y_i is generated for observation i. 1000 samples, each of size $n = 1000$, are generated. The NB, ZINB and EVZINB models are fitted to each sample. The true parameter values are set as $\boldsymbol{\gamma}_Z^T = (0.5, -0.3, 0.3, -0.3)$, $\boldsymbol{\gamma}_{EV}^T = (-1.0, 0.3, -0.3, 0.3)$ so that approximately 54% of the observations are generated from the Z state, 36% from the NB state and 14% from the EV state, $\boldsymbol{\beta}_{NB}^T = (3.0, 0.2, 0.2, 0.2)$, $\alpha_{NB} = 1.0$ so that the mean of the NB process fluctuates around 25, representing a moderate count process, with substantial overdispersion. $\boldsymbol{\beta}_{EV}^T = (0.80, 0.05, 0.05)$ is set so that the shape parameters $\alpha_{EV,i}$ fluctuates around 2 with a small fraction less than 1, representing potentially highly extreme values, and $c_{EV} = 150$ as the lower limit for the EV process. These data, hence, represents a plausible empirical distribution with a large fraction of zeroes and a few extreme observations, generated from the EV state.[9] All covariates are used for modeling $\boldsymbol{\gamma}_Z$, $\boldsymbol{\gamma}_{EV}$, $\boldsymbol{\beta}_{NB}$ and $\boldsymbol{\beta}_{EV}$.

In order to test the sensitivity of the models to the magnitude of the EV state, three additional sets of parameters are investigated. In the first set $\boldsymbol{\gamma}_{EV}$ is chosen so that approximately 1.9% of the observations are generated from the EV state. In the second set $\boldsymbol{\gamma}_{EV}$ is chosen so that approximately 0.7% of the observations are generated from the EV state. In the third set no EV state is present. In the third case we expect the EVZINB model assuming no EV state is equivalent to the ZINB model.

Simulation results

Tables 3, 4 and 5 show population value, average estimate, bias, SD and RMSE of parameters estimated using EVZINB, ZINB and NB models, respectively. Only the the proposed EVZINB model provides parameter estimates with low bias. The RMSLE is dominated by the variance. Using ZINB and NB, all parameters are associated with substantial bias which constitutes a large proportion of the RMSLE. Observe in particular that the average estimate of the negative binomial dispersion parameter α_{NB} is approximately 2.78 for ZINB and 9.27 for NB instead of its true value of 1. This discrepancy is an effect of the fact that ZINB needs to account for extreme values using the negative binomial (NB) state which leads to an enhancement of the conditional variance. This effect is further enhanced in NB because then also the excess zeroes need to be accounted for using α_{NB}.

Table 3. Population value, average estimate, bias, standard deviation (SD) and root mean squared error (RMSE) of model parameters from 1000 simulated samples using EVZINB. There are no inadmissible cases.

	Population value	Average estimate	Bias	SD	RMSE
$\gamma_{Z,0}$	0.5	0.502	0.002	0.080	0.080
$\gamma_{Z,1}$	-0.3	-0.309	-0.009	0.089	0.089
$\gamma_{Z,2}$	0.3	0.303	0.003	0.085	0.085
$\gamma_{Z,3}$	-0.3	-0.299	0.001	0.084	0.084
$\gamma_{EV,0}$	-1	-1.011	-0.011	0.118	0.118
$\gamma_{EV,1}$	0.3	0.301	0.001	0.128	0.128
$\gamma_{EV,2}$	-0.3	-0.304	-0.004	0.121	0.121
$\gamma_{EV,3}$	0.3	0.301	0.001	0.115	0.115
$\beta_{NB,0}$	3	2.993	-0.007	0.063	0.063
$\beta_{NB,1}$	0.2	0.198	-0.002	0.075	0.075
$\beta_{NB,2}$	0.2	0.204	0.004	0.071	0.071
$\beta_{NB,3}$	0.2	0.203	0.003	0.072	0.072
α_{NB}	1	0.995	-0.006	0.120	0.120
$\beta_{P,0}$	0.8	0.823	0.023	0.102	0.105
$\beta_{P,1}$	-0.1	-0.103	-0.003	0.099	0.099
$\beta_{P,2}$	-0.1	-0.098	0.002	0.097	0.097
$\beta_{P,3}$	-0.1	-0.097	0.003	0.095	0.095
c_{EV}	150	150.480	0.480	0.820	0.950

[9] This distribution of the dependent variable resembles the distribution of the dependent variable encountered in the empirical example in section 4.

Table 4. Population value, average estimate, bias, standard deviation (SD) and root mean squared error (RMSE) of model parameters from 1000 simulated samples using ZINB. There are no inadmissible cases.

	Population value	Average estimate	Bias	SD	RMSE
$\gamma_{Z,0}$	0.5	-0.076	-0.576	0.1132	0.587
$\gamma_{Z,1}$	-0.3	-0.431	-0.131	0.0955	0.162
$\gamma_{Z,2}$	0.3	0.420	0.120	0.0921	0.151
$\gamma_{Z,3}$	-0.3	-0.424	-0.124	0.0890	0.152
$\beta_{NB,0}$	3	4.358	1.3578	0.097	1.361
$\beta_{NB,1}$	0.2	0.273	0.0734	0.119	0.140
$\beta_{NB,2}$	0.2	-0.078	-0.2776	0.111	0.299
$\beta_{NB,3}$	0.2	0.269	0.0690	0.106	0.126
α_{NB}	1	2.777	1.7768	0.380	1.817

Table 5. Population value, average estimate, bias, standard deviation (SD) and root mean squared error (RMSE) of model parameters from 1000 simulated samples using NB. There are 9 inadmissible cases.

	Population value	Average estimate	Bias	SD	RMSE
$\beta_{NB,0}$	3	3.6509	0.651	0.106	0.524
$\beta_{NB,1}$	0.2	0.4843	0.284	0.128	0.289
$\beta_{NB,2}$	0.2	-0.2858	-0.486	0.121	0.376
$\beta_{NB,3}$	0.2	0.4712	0.271	0.112	0.287
α_{NB}	1	9.2745	8.275	0.545	8.373

Table 6 shows standard deviations of $\boldsymbol{\beta}_{NB}$ and α_{NB} from the simulations. Parameters occurring in all three methods (EVZINB, ZINB and NB) are included for comparison. The smallest standard deviation is obtained using EVZINB for all five parameters, indicating that parameter estimates using EVZINB are more efficient than those using ZINB and NB.

Table 6. Standard deviation (SD) of parameter estimates from simulationed samples estimated by EVZINB, ZINB and NB, respectively.

	EVZINB	ZINB	NB
$\beta_{NB,0}$	0.063	0.102	0.130
$\beta_{NB,1}$	0.069	0.139	0.137
$\beta_{NB,2}$	0.067	0.126	0.129
$\beta_{NB,3}$	0.063	0.127	0.132
α_{NB}	0.116	0.528	0.632

Appendix B1: Regression Results

Table 7. Regression results from the Negative Binomial and Zero Inflated Negative Binomial Regressions (Moore's original model, with log of lagged fatalities instead of counts)

	Negative Binomial	Zero Inflated Negative Binomial	
	NB β	NB β	ZI γ
foreign_f	0.723	0.780	−0.235
	(0.298)	(0.145)	(0.222)
rebstrength	0.707	0.365	−0.293
	(0.220)	(0.119)	(0.164)
loot	0.689	0.286	−0.506
	(0.273)	(0.119)	(0.200)
territorial	−0.805	−0.237	0.920
	(0.314)	(0.181)	(0.240)
islamist_nsa	0.681	−0.474	−0.778
	(0.944)	(0.337)	(0.708)
leftist	−0.754	0.060	0.669
	(0.401)	(0.203)	(0.315)
length	0.071	−0.011	−0.047
	(0.026)	(0.014)	(0.020)
pop_dens_ln	0.349	0.030	−0.404
	(0.124)	(0.064)	(0.097)
gdp_cap_gr	0.010	−0.004	−0.009
	(0.014)	(0.009)	(0.011)
govtbestfatal_ln	0.097	0.083	0.028
	(0.056)	(0.029)	(0.043)
log(fatalities_lag)	0.347	0.119	−0.418
	(0.063)	(0.024)	(0.043)
intensity	1.856	1.065	−1.609
	(0.422)	(0.142)	(0.312)
Constant	−1.164	3.286	4.160
	(0.655)	(0.366)	(0.526)
Observations	825	825	
Log Likelihood	−2,111.926	-1,858.727	
α_{nb}	13.230	1.348	
Akaike Inf. Crit.	4,249.852	3,771.454	
Note:		Standard errors in parenthesis	

Table 8. Regression Results from the Extreme Value and Zero Inflated Negative Binomial Regression

	Extreme Value and Zero Inflated Negative Binomial			
	NB β	ZI γ	EVI γ	Pareto β
foreign_f	0.053	−0.111	0.786	−0.495
	(0.090)	(0.247)	(0.439)	(0.187)
rebstrength	0.117	−0.187	0.514	
	(0.108)	(0.185)	(0.319)	
loot	0.112	−0.434	0.336	
	(0.101)	(0.222)	(0.450)	
territorial	0.085	0.774	−1.091	0.757
	(0.127)	(0.266)	(0.716)	(0.188)
islamist_nsa	0.115	−0.775		
	(0.344)	(0.923)		
leftist	−0.031	0.772	0.373	
	(0.150)	(0.353)	(0.643)	
length	−0.015	−0.042	0.038	
	(0.010)	(0.021)	(0.049)	
pop_dens_ln	0.024	−0.365	0.279	
	(0.043)	(0.114)	(0.209)	
gdp_cap_gr	0.004	−0.015	−0.017	
	(0.006)	(0.013)	(0.024)	
govtbestfatal_ln	0.041	0.022	−0.053	
	(0.021)	(0.048)	(0.081)	
log(fatalities_lag)	0.025	−0.333	0.414	−0.067
	(0.020)	(0.050)	(0.106)	(0.037)
intensity	0.057	−1.139	1.859	
	(0.165)	(0.409)	(0.533)	
Constant	3.644	3.895	−5.150	0.442
	(0.277)			(0.253)
Observations			825	
Log Likelihood			-1,794.011	
α_{nb}	0.127			
C_{EV}				123.274
Akaike Inf. Crit.			3,676.023	

Note: Standard errors in parenthesis

Likelihood ratio test

Table 9 displays LR tests of all covariates. Degrees of freedom (ν), the value of corresponding statistic and p-value is presented for each covariate.

Table 9. LR test of all covariates. Degrees of freedom (ν), value of statistic and p-value are presented.

variable	ν	statistic	p
foreign_f	4	10.658	0.031
rebstrength	3	8.372	0.039
loot	3	9.520	0.023
territorial	4	28.632	0.000
islamist_nsa	2	1.200	0.549
leftist	3	5.191	0.158
length	3	12.307	0.006
pop_dens_ln	3	20.064	0.000
gdp_cap_gr	3	2.508	0.474
govtbestfatal_ln	3	5.763	0.124
fatality_lag_ln	4	138.067	0.000
intensity	3	49.557	0.000

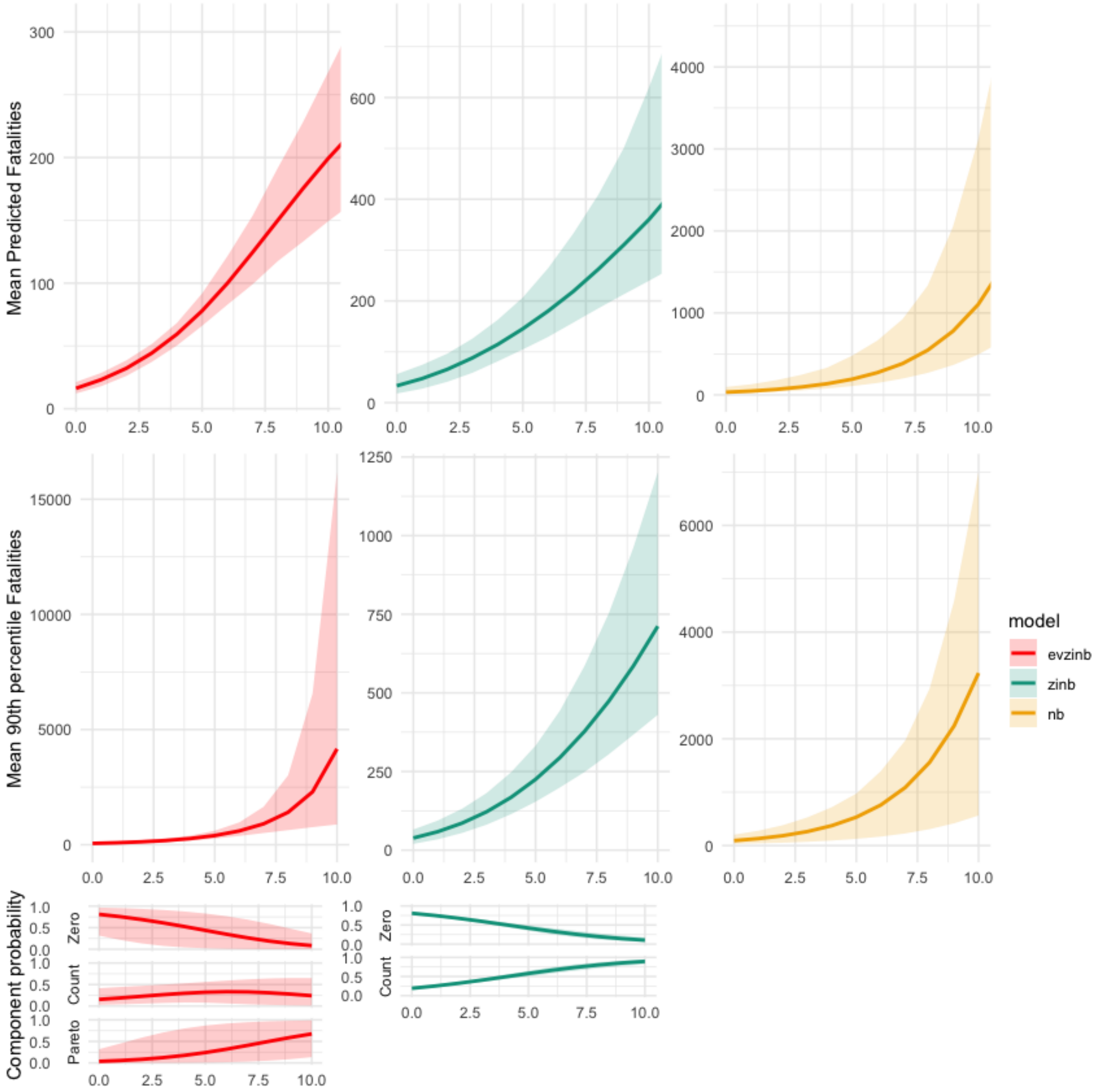

Figure 4. Mean predicted fatalities, mean 90th percentile, and state probabilities for different values of *log(fatalities_lag)* across all bootstrapped samples with the remaining covariates set to their observed value, with 95% bootstrapped intervals.

Appendix B2: Model performance metrics

Model fit

Table 10. Bootstrapped† paired difference in AIC and BIC between models and proportion of bootstrapped samples with EVZINB AIC and BIC lower than compared model. Standard errors in parenthesis

Comparison, M1 vs M2	Mean diff AIC	Mean diff BIC	Prop M1<M2 AIC	Prop M1<M2 BIC
EVZINB vs NB	-616.930	-475.469	1	1
	(1.766)	(1.766)	(0.000)	(0.000)
EVZINB vs ZINB	-107.7677	-27.606	0.997	0.750
	(1.343)	(1.343)	(0.002)	(0.014)

Note: Standard errors in parenthesis

†483 bootstraps for NB and 999 bootstraps for ZINB. 15 bootstraps with extreme values were excluded for NB, as were 502 inestimable bootstraps for NB and 1 inestimable bootstrap for ZINB. No bootstrapped samples were inestimable for EVZINB. Excluding extreme values and inestimable bootstrapped samples strictly benefits the NB and ZINB models in the comparison.

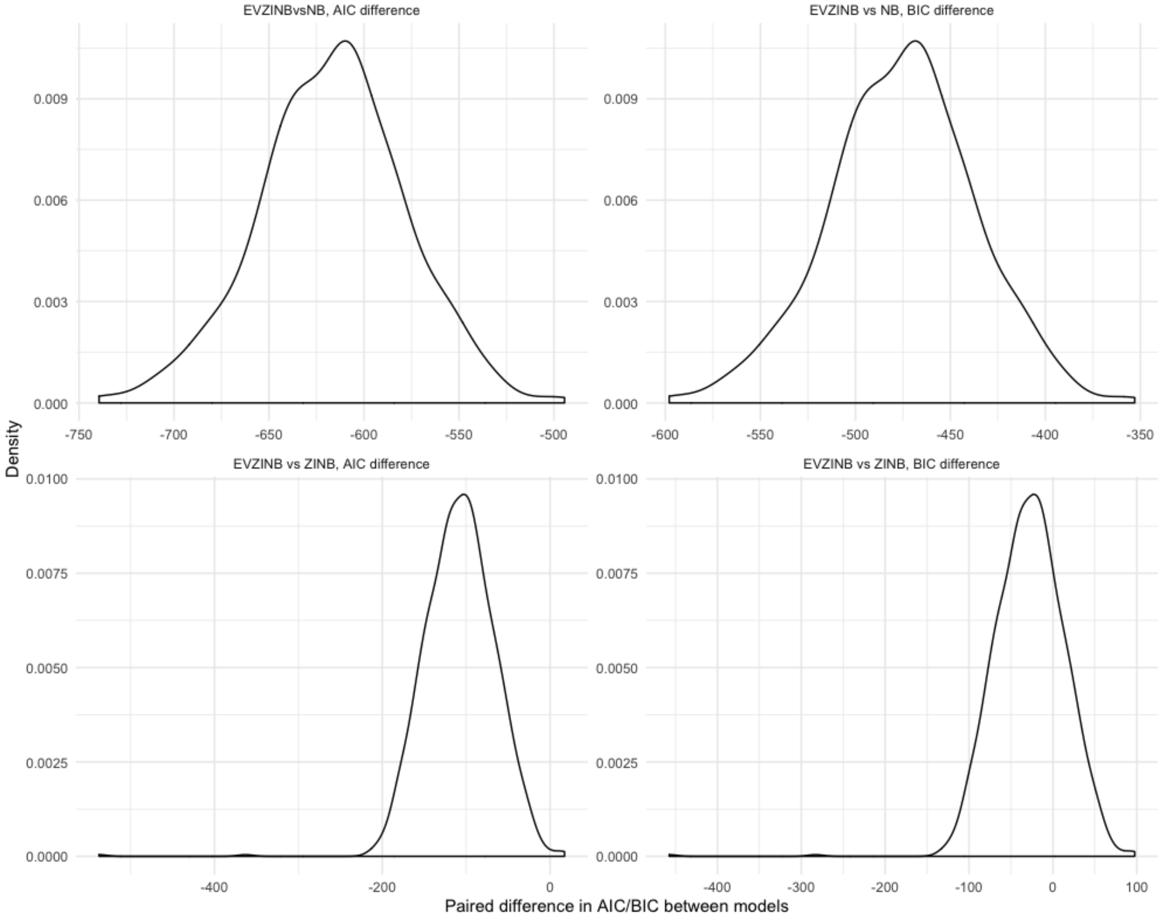

Figure 5. Densities of bootstrapped paired difference in AIC and BIC between models