SORT 41 (1) January-June 2017, 177-188

Statistical modelling of warm-spell duration series using hurdle models

Jesper Rydén*

Abstract

Regression models for counts could be applied to the earth sciences, for instance when studying trends of extremes of climatological quantities. Hurdle models are modified count models which can be regarded as mixtures of distributions. In this paper, hurdle models are applied to model the sums of lengths of periods of high temperatures. A modification to the common versions presented in the literature is presented, as left truncation as well as a particular treatment of zeros is needed for the problem. The outcome of the model is compared to those of simpler count models.

MSC: 62J12, 62P12.

Keywords: Count data, hurdle models, Poisson regression, negative binomial distribution, climate.

1. Introduction

Regression models for counts arise when the response variable is a count, i.e. a non-negative random number. Often a distribution is specified for the response variable and likelihood-based inference can be performed, with maybe the most common choice for the response being the Poisson distribution, leading to Poisson regression. However, the simpler models are not able to successfully model situations with, for example, excess zeros or truncated observations. Models have been developed in the literature, see Cameron and Trivedi (2013) for a review.

Some statistical problems in the earth sciences are linked to count data. In particular, regression models for count data could be of interest, when facing series of so-called climate indicators (occasionally called climate indices). These are often numbers relating to extreme phenomena, for instance *heat waves* or *warm spells*, loosely described

* Department of Mathematics. Uppsala University. Box 480. SE 751 06 Uppsala. Sweden.
Office phone: +46 18 4713288. E mail: jesper.ryden@math.uu.se
Received: March 2016
Accepted: February 2017

as periods of unusually hot weather. In the literature, the notion of a heat wave is often reserved for periods of great severity, for instance causing deaths among people. For typical Swedish conditions, analysed in the sequel, the notion warm spell is therefore preferred. In climatology, interest concerns changes in frequency, intensity or duration of such quantities.

From a data-analytic point of view, a warm spell is a *run*, i.e. a period of consecutive days when the maximum is above a specified high value. In this paper, we examine statistical modelling of the indicator *warm-spell duration index* (WSDI), defined as the annual count of days with at least 6 consecutive days when the daily maximum temperature is exceeding a predefined threshold (see exact definition in the sequel). However, the statistical modelling of such sequences imply several challenges. In this paper, we focus on the fact that observations are truncated, but in addition, an annual count of zero might also be observed depending on the location and its occasionally cold climate. In fact, such models for count data seem not to have been studied in the applied literature, either in climate research or other applications.

A key issue in climatology research is investigation of trends. A methodology for count data could be to check independence, and if possible, use time as a covariate in a regression model. Generalised linear models and their extensions are then natural candidates for modelling. As stated by Chandler and Scott (2011), applications of such models in environmental trend analysis have so far been relatively limited. Examples are rare, but similar statistical concepts are found for instance in Frei and Schär (2001), a recent study on extreme precipitation was made by Hertig et al (2014), and trends of flash counts are discussed by Bates, Chandler and Dowdy (2015). Concerning an indicator related to the annual number of warm spells, Rydén (2015) investigated a possible trend for the city of Uppsala, Sweden. Then the elements of the time series were simply non-negative integers, and the Poisson distribution was found to be a reasonable description. Moreover, the sequence was considered independent, and hence Poisson regression was applied.

The paper is organised as follows. In the next section, the indicator WSDI is defined and discussed, along with a presentation of the source of the data. In Section 3, the framework of hurdle models is introduced, including the modification needed for modelling of the WSDI. In Section 4, data are introduced and the results of applying the hurdle models are presented, and finally in Section 5, a summary and discussion is given.

2. Warm-spell duration index

Several indicators, also labelled indices, have been suggested for monitoring change in climatic extremes (see e.g. Frich et al 2002), but as pointed out by Perkins and Alexander (2013) concerning heat waves: “Clear and common definitions, at least for some

types of extreme events, remain rare and nonexistent". Climate indices may relate to temperatures as well as precipitation, they may be based on absolute thresholds or percentile based. Thus, definitions have to be clearly stated in research work. An overview of indices, as well as results from an analysis of trends at a global level, is given by Alexander et al (2006).

Data were retrieved online from the website of the European Climate Assessment & Dataset (ECA&D) project[1]. Definitions of indices are found at the webpage of the joint CCl/CLIVAR/JCOMM Expert Team (ET) on Climate Change Detection and Indices (ETCCDI)[2].

The indicator WSDI, warm-spell duration index, belongs to the category of duration indices. Such indices define periods of excessive warmth, cold, wetness or dryness. WSDI is defined as the annual count of days with at least 6 consecutive days when the daily maximum temperature is exceeding the threshold T_{90}. To be more precise: Let $T(i,j)$ be the daily maximum temperature on day i in year j and let T_{90} be the calendar day 90th percentile, centred on a five-day window for the base period 1961-1990. Then the number of days per year j is summed where, in intervals of at least 6 consecutive days, $T(i,j) > T_{90}$.

Note that the annual count, the annual observation of WSDI is a sum of all days belonging to a warm-spell period. The number of warm spells is not taken into account, so a year with two spells of lengths 6 and 8 days, respectively, would result in a value of WSDI equal to 14, the same value as a year with a single long spell of 14 days.

3. Hurdle models for count data

In this section we review hurdle models for count data (cf. Winkelmann 2008, Cameron and Trivedi 2013), and discuss implications for the application introduced previously and possible alternatives for the modelling.

3.1. Structure of the hurdle-count model

We commence by recalling the notion of a truncated random variable. Consider a random variable Y, defined on $0,1,2,\ldots$. Now assume that only values $y > a$ are observed. The truncated distribution $\widetilde{Y}$ then has the probability-mass function

$$p_{\widetilde{Y}}(\widetilde{y}) = \frac{1}{1-F_Y(a)} p_Y(\widetilde{y}), \quad \widetilde{y} = a+1, a+2, \ldots \ .$$

1. http://www.ecad.eu/
2. http://etccdi.pacificclimate.org/list_27_indices.shtml

With two-part models for counts, a model is introduced where the probabilistic properties of zero counts differ from other (positive) counts. Such models were proposed by Mullahy (1986). For a random variable Y, suppose that we observe either $Y = 0$ or $Y > a$. For the zero component, we introduce the probability-mass function $p_1(y)$ and for the positive outcomes, we consider the (unrestricted) probability-mass function $p_2(y)$; related distribution functions are $F_1(y)$ and $F_2(y)$. A *hurdle model* is then defined by

$$\mathsf{P}(Y = j) = \begin{cases} p_1(0) & \text{if } j = 0 \\ \frac{1-p_1(0)}{1-F_2(a)} p_2(j) & \text{if } j > a \end{cases} \tag{1}$$

(For $0 < j \leq a$, the probability-mass function takes the value zero.) Defining a binary, censoring indicator

$$d = \begin{cases} 1, & \text{if } y > a \\ 0, & \text{if } y = 0 \end{cases}$$

the probability-mass function for an outcome y with indicator d can then be written as

$$\begin{aligned} p(y) &= p_1(0)^{1-d} \left[\frac{1-p_1(0)}{1-F_2(a)} p_2(y) \right]^d \\ &= \left[p_1(0)^{1-d} (1-p_1(0))^d \right] \left[\frac{p_2(y)}{1-F_2(a)} \right]^d . \end{aligned}$$

3.2. Estimation

We now turn to estimation. In a regression context, suppose we have a covariate x. Introducing parameter vectors $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, the probability functions can be notated $p_1(y; x, \boldsymbol{\theta}_1)$ and $p_2(y; x, \boldsymbol{\theta}_2)$.

The log-likelihood function then follows, with observations $(x_1, y_1), \ldots, (x_n, y_n)$, as

$$\begin{aligned} \ell(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) &= \ell_1(\boldsymbol{\theta}_1) + \ell_2(\boldsymbol{\theta}_2) \\ &= \sum_{i=1}^{n} \left[(1-d_i) \ln p_1(0; x_i, \boldsymbol{\theta}_1) + d_i \ln(1 - p_1(0; x_i, \boldsymbol{\theta}_1)) \right] \\ &\quad + \sum_{i=1}^{n} d_i \left[\ln p_2(y_i; x_i, \boldsymbol{\theta}_2) - \ln(1 - F_2(a; x_i, \boldsymbol{\theta}_2)) \right]. \end{aligned} \tag{2}$$

Thus, the log-likelihood function can be maximised by separately maximising each component, which certainly simplifies the numerical treatment.

3.3. Specification of count distributions

Many options obviously exist for choosing the distributions $p_1(.)$ and $p_2(.)$. In the original paper by Mullahy (1986), these were specified to be of the same family. Common practice now is to specify different processes for $p_1(.)$ and $p_2(.)$. The binary process, $p_1(.)$, is often modelled as a logit model, while $p_2(.)$ is chosen as a Poisson or negative binomial distribution. After preliminary analysis of data, overdispersion was found present and truly significant (p value $1.6 \cdot 10^{-4}$, test by Cameron and Trivedi (1990), as implemented in the routine `dispersiontest` in the R package `AER`, see Kleiber and Zeileis, 2008). Thus a negative binomial distribution was applied, and will be discussed next.

Several characterisations of the negative binomial distribution exist, in terms of parameterisation, and we chose in this work to employ the distribution with probability-mass function as follows, the so-called Negbin II:

$$p(z;\mu,\alpha) = \frac{\Gamma(\alpha^{-1}+z)}{\Gamma(\alpha^{-1})\Gamma(z+1)} \left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu}\right)^{1/\alpha} \left(\frac{\mu}{\alpha^{-1}+\mu}\right)^{z}, \quad z = 0,1,2,\ldots \qquad (3)$$

where in a regression context with a covariate x,

$$\mu = \exp(\beta_0 + \beta_1 x)$$

and α is a dispersion parameter. When $\alpha \to 0$, the Poisson distribution is obtained as a limit. Moreover,

$$\mathsf{E}[Z] = \mu, \qquad \mathsf{V}[Z] = \mu(1+\alpha\mu).$$

In summary; with regards to the likelihood estimation in our problem, we have the parameter vector $\boldsymbol{\theta}_2 = (\beta_0,\ \beta_1,\ \alpha)$. For the binary part,

$$p_1(y, x_i, \boldsymbol{\theta}_1) = \frac{\exp(\beta_0' + \beta_1' x_i)}{1+\exp(\beta_0' + \beta_1' x_i)}$$

and thus $\boldsymbol{\theta}_1 = (\beta_0',\ \beta_1')$.

Remark. In most texts, and computer implementations (in R, e.g. Zeileis, Kleiber and Jackman, 2008), hurdle models with $a = 0$ in Eq. (1) are considered; that is, the hurdle separates zeros from positive observations. In our application, we have the possible outcomes $0, 6, 7, 8, \ldots$, and hence $a = 5$ in Eq. (1); to the author's knowledge, this is a situation rarely met in applications considered in the literature. In Stata, modelling with truncated hurdle models is implemented. An example with a truncated Poisson distribution for the non-zero counts is given by McDowell (2003).

3.4. Mean for the hurdle-count model

For the hurdle model in Eq. (1), for the sake of notation, introduce

$$b = \frac{1 - p_1(0)}{1 - F_2(a)}. \tag{4}$$

Moments about the origin then follow as

$$\mathsf{E}[Y^k] = 0^k p_1(0) + \sum_{y=a+1}^{\infty} y^k b\, p_2(y) = b \sum_{y=a+1}^{\infty} y^k p_2(y). \tag{5}$$

Cameron and Trivedi (2013), Section 4.12, give the corresponding derivation for the case where $a = 0$, and the resulting formula can then be expressed in terms of the expected value for the distribution $F_2(y)$.

For model diagnostics (in our application, see Section 4.3), we may use $\mathsf{E}[Y]$ following Eq. (5), plugging in estimates. Then assuming a logit model for the binary part, we find estimates for the quantities in the factor b in Eq. (4):

$$\widehat{p_1(0)} = p_1(0; x_i, \widehat{\boldsymbol{\theta}}_1) = \frac{\exp(\widehat{\beta}_0' + \widehat{\beta}_1' x_i)}{1 + \exp(\widehat{\beta}_0' + \widehat{\beta}_1' x_i)}$$

and $\widehat{F_2(a)} = F_2(a, x_i, \widehat{\boldsymbol{\theta}}_2)$, where $F_2(.)$ is the related distribution for the Negbin II distribution.

4. Modelling warm-spell duration index

We have chosen to investigate time series of annual observations of WSDI from three locations and periods in Sweden: Falun (60°37′N, 15°37′E), 1914-2010, Stockholm (59°21′N, 18°03′E), 1914-2014, and Uppsala (59°51′N, 17°37′E), 1914-2011. This initial choice was made based on data quality (quite long series without gaps, though missing data for 2006 at Falun) and we have, moreover, two locations quite close in distance (Stockholm and Uppsala, less than 10 km).

4.1. Dependence issues

In order to apply regression models for counts using time as a covariate, dependence in each of the sequences was first investigated, by checking plots of autocorrelation functions and performing the Ljung–Box test of independence in time series (Ljung and Box, 1978). Consider a time series $x_1, \ldots, x_N$. The null hypothesis is here that the first

m autocorrelations are jointly zero:

$$H_0: \quad \rho_1 = \rho_2 = \cdots = \rho_m = 0.$$

Let r_k be the sample autocorrelation function at lag k and m the number of lags being tested. The test statistic is given by

$$Q = N(N+2)\sum_{k=1}^{m} \frac{r_k^2}{N-k}$$

which under the null hypothesis is distributed as $Q \sim \chi^2(m)$. A choice of m has to be made; in the literature, it has been suggested that $m \approx \ln N$ (Tsay, 2010). For our locations we find Falun ($p = 0.85$), Stockholm ($p = 4.3 \cdot 10^{-10}$), Uppsala ($p = 6.2 \cdot 10^{-8}$). For the two last locations, we thus reject the null hypothesis about independence. Plots of empirical autocorrelation functions strengthen this result. Thus, a more evolved time-series model for counts would have to be introduced for these two locations. One option for further modelling could be to introduce a time-series model for counts, for instance, of the class INAR (Al-Osh and Alzaid, 1987; Alzaid and Al-Osh, 1990; for a recent review, see Scotto, Weiss and Gouveia, 2015).

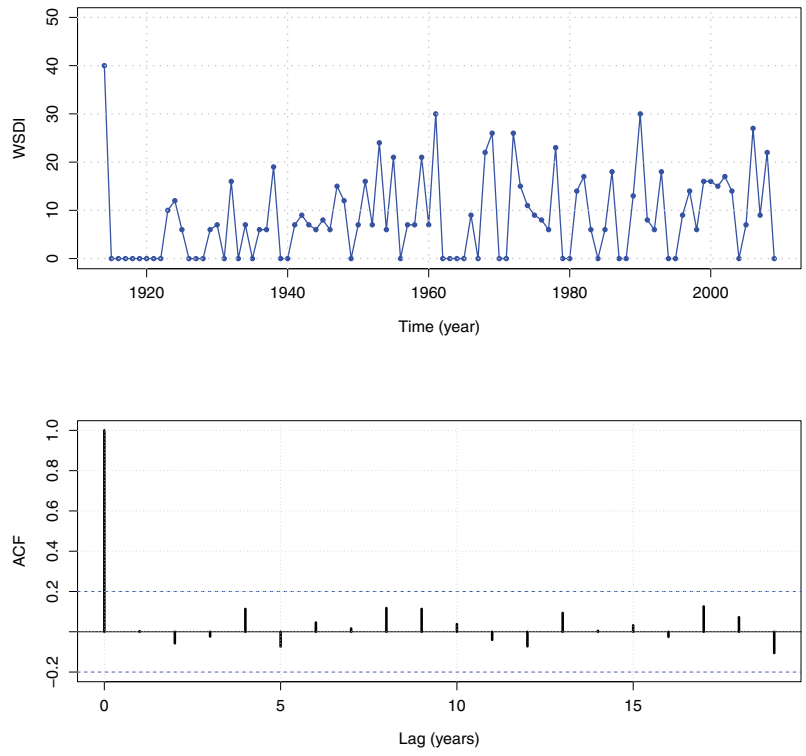

Figure 1: *Top: Time series of WSDI at Falun. Bottom: Sample autocorrelation function of WSDI at Falun.*

For the location of Falun, we choose to continue, using time as a covariate in a regression model. The original time series[3] at that location is shown in Figure 1, top panel. The proportion of zeros is found to be 0.35. The sample autocorrelation function is displayed in Figure 1, bottom panel, and we conclude that data could be considered a sequence of independent observations.

4.2. Likelihood inference

For the zero part, i.e. maximisation of the function $\ell_1(\boldsymbol{\theta}_1)$ in Eq. (2), a logistic regression was performed with a binary response variable $\pi(x)$ and the related model

$$g(x) = \beta_0' + \beta_1' x$$

where $g(x) = \ln(\pi(x)/(1-\pi(x)))$ and the covariate x indicates time. Hence, the vector $\boldsymbol{\theta}_1 = (\beta_0' \ \beta_1')$. The routine `glm` in the statistical software package R (R Core Team 2016) was employed. The estimation procedure resulted in estimates $\hat{\beta}_0' = -0.25$ and $\hat{\beta}_1' = 0.018$ with related p-values 0.56 and 0.025, respectively. The covariate time is thus significant.

For the maximisation of the log-likelihood function $\ell_2(\boldsymbol{\theta}_2)$, a Negbin II was assumed (see Eq. (3)). The optimisation was carried out by the routine `optim`, using the procedure by Nelder and Mead (1965). The following point estimates, with related standard errors within parentheses as obtained from the inverted observed Fisher information matrix, were obtained:

$$\hat{\beta}_0 = 1.01 \ (1.70), \quad \hat{\beta}_1 = 0.0053 \ (0.0062), \quad \alpha = 2.84 \ (5.98)$$

with p-values 0.56, 0.39 and 0.64 respectively. With the climate application in focus, we note that the slope is slightly positive in magnitude, and not statistically significant.

4.3. Model checking and comparison

In Figure 2, the original time series is plotted along with the mean of the fitted model, following Eq. (5).

We deduced earlier that the original time series could be considered an independent sequence (cf. Figure 1, bottom panel). In Figure 3, the sample autocorrelation function of the raw residuals is shown. We note that dependence is still not a concern.

3. The high observed value 40 at the beginning of the series belongs to year 1914, the summer of which is known for historians as being unusually warm and pleasant, right before the outburst of World War I, end of July 1914.

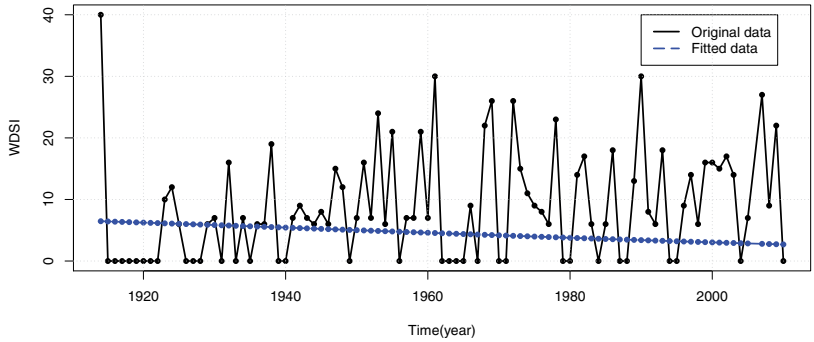

Figure 2: *Original time series and fitted model.*

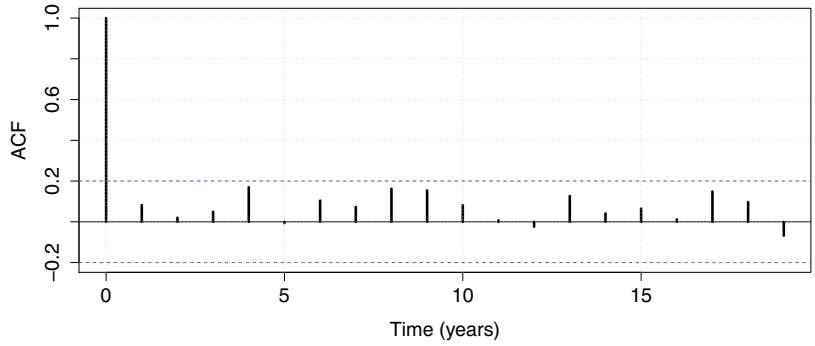

Figure 3: *Sample autocorrelation function for the residuals.*

One might contemplate a simpler statistical model, though not taking into account the particular structure of data. A negative binomial distribution with mean $\mu = \exp(\beta_0 + \beta_1 x)$, say, could then be directly fitted to all outcomes of WSDI in a regression model (using the routine `glm.nb` in the package MASS in R). Such a model results in an estimate $\hat{\beta}_1 = 0.0092$ for the slope (standard error 0.0055) with the related p-value 0.092, which could be compared to the corresponding estimate for the hurdle model.

For model comparison, values of AIC (Akaike's Information Criterion, Akaike 1973) are useful:

$$\mathrm{AIC} = 2k - 2\ln L$$

where k is the number of parameters and L the value at the optimum of the likelihood function. These were computed for the two considered models. For the hurdle model, AIC $= 505.7$, while for the approach with negative binomial, AIC $= 597.8$. A model with as small AIC as possible is preferable, and there is hence some merit of the hurdle model in this respect.

5. Discussion

Regression models for counts find applications in many scientific fields. Typically, a Poisson distribution is assumed for count data, but the original models have to be modified in order to model e.g. overdispersion or excesses of zeros. One special model is the so-called hurdle model, attributed to Mullahy (1986). As a further, quite recent, development of the hurdle model could be mentioned Saffari, Adnan and Greene (2012), where a framework with hurdle models adopted to right-censored data was presented (application to counts of fish).

In this paper, a hurdle model for the case of left-truncated data was presented, motivated by an application from climatology where data is either zero or an integer at least six. Estimation was carried out by likelihood techniques. The obtained results were compared with estimates from a simpler model, fitting a negative binomial distribution directly to the counts. Point estimates of trend (coefficient for slope) became of roughly the same magnitude. Comparison of AIC indicates that the hurdle model is preferable. However, for all estimated parameters, uncertainties are considerably high, as can be reflected from related p-values.

The meaning of the quantity WSDI as an additive measure of days may have influences on the distribution over possible integers. For instance, a WSDI of 11 can be obtained only as a single period of 11 days, while an observed count of 14 can result in three ways: a single period of 14, adding 6 and 8 or adding 7 and 7. Thus, in addition to natural variability, results could vary due to combinatoric reasons. For the data sets, the following table of counts for various WSDI can be compiled (for Falun, also cf. Figure 1, top panel):

WSDI	7	8	9	10	11	12	13	14	15
Counts, Falun	10	3	5	1	1	2	1	3	3
Counts, Uppsala	6	4	3	1	2	4	6	6	2
Count, Stockholm	6	5	4	1	0	2	3	2	4

We note that for WSDI equal to 10 and 11, few counts are found (not likely combinations to occur). Thus, to model WSDI with a probability distribution, possibly this phenomenon could be taken into account.

For all types of regression models, model assessment is an important objective. A review for the common cases of regression with count data is given by Cameron and Trivedi (2013), where it is also stated in Chapter 5 (Model Validation and Testing) that there is "... considerable scope for generalization and application to a broader range of count data models." In this paper, we made a simple investigation (see Figures 2 and 3). Further research would be to, for instance, develop and examine goodness-of-fit tests for the hurdle model with truncated observations for the non-zero part.

In this paper, regression models for counts were modelled using time as a covariate. It could be mentioned that a non-parametric regression methodology based on P-splines

might be a useful approach, see the recent paper by Eilers, Marx and Durbán (2015) in this journal.

Acknowledgements

I am grateful to Dr Hans Bergström, Dept. of Earth Sciences at Uppsala University, for discussion about climate indicators. Thanks also to the referees, whose constructive comments led to improvement of the paper.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In: Petro, B.V. and Csaki,F., eds., *Second International Symposium on Information Theory*, 267-281, Budapest, Akademiai Kaido.

Alexander, L.V., Zhang, X., Peterson, T.C., Caesar, J., Gleason, B., Klein Tank, A.M.G., Haylock, M., Collins, D., Trewin, B., Rahimzadeh, F., Tagipour, A., Ambenje, P., Rupa Kumar, K., Revadekar, J. and Griffiths. G. (2006). Global observed changes in daily climate extremes of temperature and precipitation. *Journal of Geophysical Research*, 111, 1–22.

Al-Osh, M.A. and Alzaid, A.A. (1987). First order integer valued autoregressive (INAR(1)) process. *Journal of Time Series Analysis*, 8, 261–275.

Alzaid, A.A. and Al-Osh, M. (1990). An integer-valued pth-order autoregressive structure (INAR(p)) process. *Journal of Applied Probability*, 27, 314–324.

Bates, B.C., Chandler, R.E. and Dowdy, A.J. (2015). Estimating trends and seasonality in Australian monthly lightning flash counts. *Journal of Geophysical Research: Atmospheres*, 120, 3973–3983.

Cameron, A.C. and Trivedi, P.K. (1990). Regression-based tests for overdispersion in the Poisson model. *Journal of Econometrics*, 46, 347-364.

Cameron, A.C. and Trivedi, P.K. (2013). *Regression Analysis of Count Data*. 2nd ed. Cambridge University Press.

Chandler, R.E. and Scott, E.M. (2011). *Statistical Methods for Trend Detection and Analysis in the Environmental Sciences*. Wiley.

Eilers, P.H.C., Marx, B.D. and Durbán, M. (2015). Twenty years of P-splines. *SORT-Statistics and Operations Research Transactions*, 39, 149–186.

Frei, C. and Schär, C. (2001). Detection probability of trends in rare events: theory and applications to heavy precipitation in the Alpine region. *Journal of Climate*, 14, 1568–1584.

Frich, P., Alexander, L.V., Della-Marta, P., Gleason, B., Haylock, M., Klein Tank, A.M.G. and Peterson, T. (2002). Observed coherent changes in climatic extremes during the second half of the twentieth century. *Climate Research*, 19, 193–212.

Hertig, E., Seubert, S., Paxian, A., Vogt, G., Paeth, H. and Jacobeit, J. (2014). Statistical modelling of extreme precipitation indices for the Mediterranean area under future climate change. *International Journal of Climatology*, 34, 1132–1156.

Kleiber, C. and Zeileis, A. (2008). *Applied Econometrics with R*. Springer-Verlag.

Ljung, G.M. and Box, G.E.P. (1978). On a measure of lack of fit in time series models. *Biometrika*, 65, 297–303.

McDowell, A. (2003). From the help desk: hurdle models. *The Stata Journal*, 3, 178–184.

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

Nelder, J.A. and Mead, R. (1965). A simplex algorithm for function minimization. *Computer Journal*, 7, 308–313.

Perkins, S.E. and Alexander L.V. (2013). On the measurement of heat waves. *Journal of Climate*, 26, 4500–4517.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www-R-project.org/

Rydén, J. (2015). A statistical analysis of trends for warm and cold spells in Uppsala by means of counts. *Geografiska Annaler: Series A, Physical Geography*, 97, 431–436.

Saffari, S.E., Adnan, R. and Greene, W. (2012). Hurdle negative binomial regression model with right censored count data. *SORT – Statistics and Operations Research Transactions*, 36, 181–194.

Scotto, M.G., Weiss, C.H. and Gouveia, S. (2015). Thinning-based models in the analysis of integer-values time series: a review. *Statistical Modeling*, 15, 590–618.

Tsay, R.S. (2010). *Analysis of Financial Time Series*. 3rd ed. Hoboken, NJ: John Wiley & Sons.

Winkelmann, R. (2008). *Econometric Analysis of Count Data*. 5th ed. Springer-Verlag

Zeileis, A., Kleiber, C. and Jackman, S. (2008). Regression models for count data in R. *Journal of Statistical Software*, 27, 1–25.