Nat. Hazards Earth Syst. Sci. Discuss., 3, 1775–1787, 2015
www.nat-hazards-earth-syst-sci-discuss.net/3/1775/2015/
doi:10.5194/nhessd-3-1775-2015

Natural Hazards and Earth System Sciences
Open Access
Discussions

Brief Communication: Statistical detection and modeling of the over-dispersion of winter storm occurrence

M. Raschke

Freelancer, Stolze-Schrey-Str. 1, Wiesbaden, Germany

Received: 1 February 2015 – Accepted: 20 February 2015 – Published: 3 March 2015

Correspondence to: M. Raschke (mathiasraschke@t-online.de)

Published by Copernicus Publications on behalf of the European Geosciences Union.

Abstract

In this communication, I discuss and improve the detection and modeling of the over-dispersion of winter storm occurrence using the example of Germany. For this purpose, the generalized Poisson distribution and information criterions for the model selection are introduced. Correct statistical model selection ensures the statistical significance of the model, including a possible over-dispersion. Moreover, I derive the relation between expectation and variance of a thinned inhomogeneous Poisson process. This is also applied to well detect the over-dispersion in winter storm data for Germany of Karremann et al. (2014).

1 Introduction

A possible over-dispersion of the occurrence of European winter storms is subject of previous researches (e.g., Mailier et al., 2006; Pinto et al., 2013; Karremann et al., 2014) and is frequently called clustering. The issue is that an over-dispersion of a sample can be caused by randomness in the sampling; the actual storm process could follow a homogeneous Poisson. This problem is not well considered in previous researches. For example, Mailier et al. (2006) state a partly under-dispersion for European winter storms. But this could be caused by an over-fit of their Poisson regression model because a few geographically allocated parameters range from negative to positive values.

In the following section, I propose the generalized Poisson distribution as model for the storm frequency, which is more flexible than the negative binomial distribution. The latter was applied by Karremann et al. (2014) to consider the over-dispersion of winter storms. Then, I derive the behaviour of expectation and variance in the thinning process in Sect. 3. Furthermore, I explain the concept of statistical model selection in Sect. 4. It provides an indirect test of statistical significance of over-dispersion. A brief summary

NHESSD

3, 1775–1787, 2015

is given in the last section. I use the data for German winter storms of Karremann et al. (2014) to demonstrate the functionality of the proposed models and methods.

2 The generalized Poisson distribution (GPD)

The GPD has been developed by Consul and Jain (1973) and is formulated for the discrete random variable $X \geq 0$ with

$$P(x) = \frac{\lambda\left(\lambda + x\theta\right)^{x-1}}{x!} e^{-\lambda - x\theta}. \quad (1)$$

It describes the occurrence probability of $X = x$ and uses the parameters $\lambda > 0$ and θ. The expectation and variance are

$$E(X) = \lambda/(1-\theta) \text{ and} \quad (2)$$

$$V(X) = \lambda/(1-\theta)^3. \quad (3)$$

There is over-dispersion if $\theta > 0$, which means $E(X) < V(X)$. Under-dispersion is valid for $\theta < 0$ and the common Poisson distribution (PD) is formulated with $\theta = 0$. The possible modeling of under-dispersion is the great advantage of the GPD in comparison to the negative binomial distribution (NBD). The latter is applied by Karremann et al. (2014) for winter storms in Germany and considers the PD only as a limit case with an infinite parameter.

The parameters of the GPD can be estimated by the well-known maximum likelihood estimation (Lindsey, 1996; Upton and Cook, 2006). Therein, the logarithmized likelihood is a function of parameter vector $\boldsymbol{\theta}$ with

$$\log(L(\boldsymbol{\theta})) = \sum_x n_x \log(P(x;\boldsymbol{\theta})) \quad (4)$$

for the discrete distributions. The parameter vector $\boldsymbol{\theta}$ with the highest value of $\log(L(\boldsymbol{\theta}))$ is the point estimation. The likelihood estimation is asymptotically the best estimation if

NHESSD
3, 1775–1787, 2015

some weak conditions are fulfilled (Cramér–Rao bound, Cramér 1946), as it applies for the GPD with over-dispersion. The ML estimation is also frequently the best estimation method for a distribution model in case of a finite sample size and is recommended for the NBD (Johnson et al., 2005) and the GPD (Consul and Shoukri, 1984). In all cases, an estimator should be bias-free and consistent and well established in mathematical statistics (Upton and Cook, 2006).

3 What does over-dispersion mean?

Mailier et al. (2006) and Karremann et al. (2014) denote the over-dispersion in the random distribution of the number of storms as serial clustering and quantify it by the following parameter

$$\phi = V(X)/E(X) - 1. \quad (5)$$

But the term cluster or clustering can mean different things. A cluster in an auto-correlated time series consists of tall observations that are a partial series of an exceedance of a threshold (Coles, 2001), e.g. of river discharge. An earthquake cluster is a group of earthquake events that include a main event with a large magnitude and a series of secondary events (after- and/or foreshocks; e.g., Ogata, 2001). There is a relation in time and space between these events. If there would be a clustering of storms similar to the clustering of earthquakes, then the number of smaller storms with return level RL = 1 and 2 should be higher for years with an event with events RL = 5. But this cannot be stated for the data of Karlemann et al. (2014, Tables 1 and B1, Supplement). It seems to be more likely that the over-dispersion is caused by an inhomogeneous intensity of the storm occurrence in time. In such an inhomogeneous Poisson process, every winter season can have different (accidental) occurrence intensity. Inhomogeneous occurrence intensity per season corresponds well with the NBD and the GPD because both are mixtures of PDs (Joe and Zhu, 2005).

NHESSD
3, 1775–1787, 2015

If an over-dispersion results from a Poisson process with inhomogeneous occurrence intensity, then an event would be independent of others. In this case, the increase of the considered RL in the sampling includes a random process of thinning. Therein, P_{survival} is the survival probability if a single event of the set of storm observations with the lower return level is also member of the set with the higher return level. For example, $P_{\text{survival}} = 1/5$ for the transition from RL ≥ 1 to RL ≥ 5. The thinning probability is $P_{\text{thining}} = 1 - P_{\text{survival}} = 4/5$. The entire sample is thinned out in this transition of RLs. According to Ross (2007, Examples 3.16 and 3.18), the expectation and variance of the new count variable X_{new} is

$$E\left(X_{\text{new}}\right) = E\left(\sum_{i=1}^{X_{\text{old}}} I_i\right) = E\left(X_{\text{old}}\right)E(I) \text{ and} \tag{6}$$

$$V\left(X_{\text{new}}\right) = V\left(\sum_{i=1}^{X_{\text{old}}} I_i\right) = E\left(X_{\text{old}}\right)V(I) + V\left(X_{\text{old}}\right)E(I)^2. \tag{7}$$

The binary random variable I describes whether the storm event i is member of the new return level ($I = 1$) or if it is thinned ($I = 0$). This random variable is Bernoulli distributed with

$$E(I) = P_{\text{survival}} \text{ and} \tag{8}$$

$$V(I) = P_{\text{survival}}\left(1 - P_{\text{survival}}\right). \tag{9}$$

Now I formulate the relation

$$V\left(X_{\text{old}}\right) = E\left(X_{\text{old}}\right) + \beta E\left(X_{\text{old}}\right)^2. \tag{10}$$

Dispersion parameter β is determined by variance and expectation of count variable X_{old}. Equation (7) is now modified for the new count variable with Eqs. (8) and (9) to

$$V\left(X_{\text{new}}\right) = E\left(X_{\text{old}}\right)P_{\text{survival}}(1 - P_{\text{survival}}) + E\left(X_{\text{old}}\right)P_{\text{survival}}^2. \tag{11}$$

NHESSD
3, 1775–1787, 2015

This equation can be simplified to

$$V\left(X_{\text{new}}\right) = E\left(X_{\text{old}}\right)P_{\text{survival}} + \beta E\left(X_{\text{old}}\right)^2 P_{\text{survival}}^2, \tag{12}$$

wherein $E(X_{\text{old}})P_{\text{survival}}$ is replaced by $E(X_{\text{new}})$ according to Eqs. (6,8) and we get

$$V\left(X_{\text{new}}\right) = E\left(X_{\text{new}}\right) + \beta E\left(X_{\text{new}}\right)^2. \tag{13}$$

Therefore, Eq. (10) is universal and also applies to the new count variables X_{new} resulting from the thinning process. This does not represent completely new knowledge and is already derived for other issues (e.g., Mack, 2002). Independent thinning is well suited to the decrease of cluster parameter ϕ with increasing return level in the sample of Karremann et al. (2014). An example of the relations is depicted in Fig. 1. Equations (10) and (13) can also be used to determine one parameter of the distributions of higher RLs (e.g., by using Eqs. (2) and (3) for the GPD).

4 How can over-dispersion be well detected?

As aforementioned, an over-dispersion of the observed sample of X and of its estimated distribution model can be caused by the randomness in the sampling. The question is whether this observed over-dispersion is statistically significant or probably a result of randomness. This could be tested by the empirical cluster parameter ϕ; sample mean and sample variance are used in Eq. (5) to estimate ϕ. It should not be smaller than the quantile of the defined exceedance probability of estimations for the Poisson case with $\phi = 0$. This can be computed by the empirical distribution (Uppton and Cook, 2006) of estimations $\hat{\phi}$ for samples for a Poisson distributed count variable X generated by Monte Carlo simulation. I have performed this calculation for sample size $n = 30$ and 10 000 repetitions. The resulting upper 5 % quantile is 0.466 for PD with $E(X) = 1$. Any value smaller than this limit means that the empirical over-dispersion is

NHESSD
3, 1775–1787, 2015

not significant for a level of $\alpha = 5\,\%$. This works similarly to a one-site t test in the regression analysis (e.g., Fahrmeir et al., 2013). For $n = 60$, the critical value is 0.323. This shows that the significance over-dispersion can be better detected for a higher sample size n.

The statistical significance of a statistical model and its parameterization can also be ensured by applying an appropriate model selection in the process of model building (e.g., Lindsey, 1996; Raschke, 2013). Information criterions are frequently used for such a selection. The Akaike information criterion (AIC, Akaike, 1973) is a very popular one and written with

$$\mathrm{AIC} = -2\log(L(\boldsymbol{\theta})) + 2m. \tag{14}$$

The Bayesian information criterion of Schwarz (1978) is written with

$$\mathrm{BIC} = -2\log(L(\boldsymbol{\theta})) + m\log(n). \tag{15}$$

The number of parameters is symbolized by m, the sample size by n. $\log(L(\boldsymbol{\theta}))$ is defined by Eq. (4) for discrete distribution. A smaller information criterion indicates the better model. The larger the difference between the criterions of alternatives is, the better the differentiation is. The AIC works better for a smaller sample size, the BIC works better for a larger sample size (Lindsey, 1996). There are further criterions, which are not considered here.

The results for the selected distribution models for samples of German winter storms (Karremann et al., 2014) are listed in Table 1. The parameters of PD, NBD and GPD are estimated by the ML method. An example of the estimated distribution is presented in Fig. 2 – the DWD sample of historical storms with a return level of one year (RL ≥ 1). Therein, NBD and GPD are very similar. Both consider over-dispersion but both distributions are not detected as the best distribution in every case of the DWD and NCEP sample of historical storms. The poor modeling of winter storm occurrence by the PD is much better detected for the large $\mathrm{GNC}_{\mathrm{corr}}$ sample of climate simulations

NHESSD
3, 1775–1787, 2015

with $n = 4092$. The over-dispersion is obvious, corresponding AICs and BICs are much smaller than for the PDs.

I have used the estimation for the GNC_{corr} sample with $RL \geq 1$ to estimate over-dispersion parameter $\hat{\beta} = 0.307$ of Eqs. (10) and (13). This relation is already shown in Fig. 1 and is used to quantify one parameter of the GPD for NCEP and DWD samples. The estimated expectation is equal to sample mean and RL, and the variance is determined by Eqs. (10) and (13). These can be transformed to the other parameter by using Eqs. (2) and (3). This estimation is possible because the GNC_{corr} sample is relatively large ($4092 \gg 30$) and independent of NCEP and DWD samples, and all samples are from the same random variable – the number of winter storms in Germany for the current climate and $RL \geq 1$. Using this special procedure with one known parameter, over-dispersion in the historic data of winter storms in Germany can very well be detected because this model is very often the best model or differs less from the best model.

I have also used Eqs. (10) and (13) with the estimation for the GNC_{corr} sample with $RL \geq 1$ for the GNC_{corr} sample with $RL \geq 2$ and 5. The resulting AICs and BICs are not correct and only a bit informative because the storms of the sample with $RL \geq 2$ and 5 are also member of the sample with $RL \geq 1$. This is not foreseen for AIC and BIC. I have tried to consider this issue by using a higher number m of parameter for Eqs. (14) and (15).

5 Summary

In this communication, I have introduced the GPD for modeling and detecting over-dispersion of winter storm occurrence because it is similar to the NBD and can also model PD and under-dispersion. Furthermore, I have derived the relation between expectation and variance in case of over-dispersion of the storm occurrence per winter season with independent events. This relation well explains the behaviour of the cluster parameter ϕ of the samples of Karremann et al. (2014) from climate situations with

NHESSD
3, 1775–1787, 2015

a large sample size. It is also very helpful for detecting over-dispersion of German winter storms in combination with the introduced criterions of model selection (AIC and BIC). The over-dispersion is likely caused by an inhomogeneous Poisson process. In contrast, a simple analysis of the sample of historical storms could not clearly detect the over-dispersion; the sample size is too small. The conclusions of Karremann et al. (2014) for winter storms in Germany have been confirmed at a higher level of statistical analysis.

The models and methods can also be applied to other hazards. Possibly also for discussing whether the same storm process with events, being independent in statistical sense, can result in over-dispersion and under-dispersion as assumed in the regression model of Mailier et al. (2006).

References

Akaike, H.: Information theory and an extension of the maximum likelihood principle, in: Proceedings of the Second International Symposium on Information Theory, edited by: Petrov, B. N., Budapest, Akademiai Kiado, Hungary, 267–281, 1973.

Coles, S.: An Introduction to Statistical Modeling of Extreme Values, Springer, London, 2001.

Consul, P. C. and Jain, G. C.: A generalization of the Poisson distribution, Technometrics, 15, 791–799, 1973.

Consul, P. C. and Shoukri, M. M.: Maximum likelihood estimation for the generalized Poisson distribution, Commun. Stat. Theory, 10, 977–991, 1984.

Cramér, H.: Mathematical Methods of Statistics, Princeton University Press, Princeton, UK, 1946.

Fahrmeir, L., Kneib, T., Lang, S., and Marx, B.: Regression – Models, Methods and Applications, Springer, Heidelberg, Germany, 2013.

Joe, H. and Zhu, R.: Generalized poisson distribution: the property of mixture of poisson and comparison with negative binomial distribution, Biometrical J., 47, 219–229, 2005.

Johnson, N. L., Kemp, A. W., and Kotz, S.: Univariate Discrete Distributions, 3rd Edn., Wiley Series in Probability and Statistics, Wiley, New York, USA, 208–247, 2005.

NHESSD
3, 1775–1787, 2015

Karremann, M. K., Pinto, J. G., von Bomhard, P. J., and Klawa, M.: On the clustering of winter storm loss events over Germany, Nat. Hazards Earth Syst. Sci., 14, 2041–2052, doi:10.5194/nhess-14-2041-2014, 2014.

Lindsey, J. K.: Parametric Statistical Inference, Oxford Science Publications, Oxford University Press, Oxford, UK, 1996.

Mack, T.: Schadenversicherungsmathematik (German; Non-Life Insurance Mathematics), 2nd Edn., Deutsche Gesellschaft für Versicherungsmathematik, Germany, 332–334, 2002.

Mailier, P. J., Stephenson, D. B., Ferro, C. A. T., and Hodges, K. I.: Serial Clustering of extratropical cyclones, Mon. Weather Rev., 134, 2224–2240, 2006.

Ogata, Y.: Exploratory analysis of earthquake clusters by likelihood-based trigger models, J. Appl. Probab., 38A, 2002–2012, 2001.

Pinto, J. G., Bellenbaum, N., Karremann, M. K., and Della-Marta, P. M.: Serial clustering of extratropical cyclones over the North Atlantic and Europe under recent and future climate conditions, J. Geophys. Res.-Atmos., 118, 12476–12485, 2013.

Raschke, M.: Statistical modelling of ground motion relations for seismic hazard analysis, J. Seismol., 17, 1157–1182, 2013.

Ross, S. M.: Introduction to Probability Models, 9th Edn., Elsevier, Amsterdam, Netherlands, 2007.

Schwarz. G.: Estimating the dimension of a model, Ann. Stat., 6, 461–464, 1978.

Upton, G. and Cook, I.: A dictionary of statistics, 2nd Edn., Oxford University Press, Oxford, UK, 2006.

NHESSD

3, 1775–1787, 2015

Statistical detection and modeling of the over-dispersion of winter storm occurrence

M. Raschke

Table 1. AIC and BIC for some samples of winter storm in Germany of Karremann et al. (2014, Tables 1 and B1), bolted: best estimation or nearly best estimation, italic: not consequently applied AIC and BIC.

Sample	Information criterion	PD, $m = 1$	NBD, $m = 2$	GPD, $m = 2$	GPD with $\beta = 0.307$ of GNC_{corr}, RL ≥ 1
GNC corrected, RL ≥ 1, $n = 4092$	AIC	11 247.24	**11 108.98**	11 110.86	–
	BIC	11 253.55	**11 121.61**	11 123.50	–
GNC corrected, RL ≥ 2, $n = 4092$	AIC	7794.04	**7745.35**	**7745.94**	*7745.63, m = 1.5*
	BIC	7800.35	**7757.99**	**7758.58**	*7755.11, m = 1.5*
GNC corrected, RL ≥ 5, $n = 4092$	AIC	4426.53	**4415.56**	**4415.52**	*4414.78, m = 1.2*
	BIC	4432.85	**4428.19**	**4428.15**	*4422.36, m = 1.2*
NCEP, RL ≥ 1, $n = 30$	AIC	**81.64**	83.57	83.57	**82.00, m = 1**
	BIC	**83.04**	86.38	86.38	**83.40, m = 1**
NCEP, RL ≥ 2, $n = 30$	AIC	**59.15**	60.77	60.78	**58.79, m = 1**
	BIC	**60.55**	63.57	63.58	**60.19, m = 1**
NCEP, RL ≥ 5, $n = 30$	AIC	36.90	**35.14**	**35.15**	**35.76, m = 1**
	BIC	38.30	**37.94**	**37.95**	**37.16, m = 1**
DWD, RL ≥ 1, $n = 30$	AIC	89.83	86.95	86.89	**85.92, m = 1**
	BIC	91.23	89.76	89.69	**87.32, m = 1**
DWD, RL ≥ 2, $n = 30$	AIC	65.14	**60.63**	**60.53**	61.81, $m = 1$
	BIC	66.54	**63.43**	**63.33**	**63.21, m = 1**
DWD, RL ≥ 5, $n = 30$	AIC	**31.30**	32.61	32.65	**31.02, m = 1**
	BIC	**32.71**	35.42	35.45	**32.42, m = 1**

NHESSD
3, 1775–1787, 2015

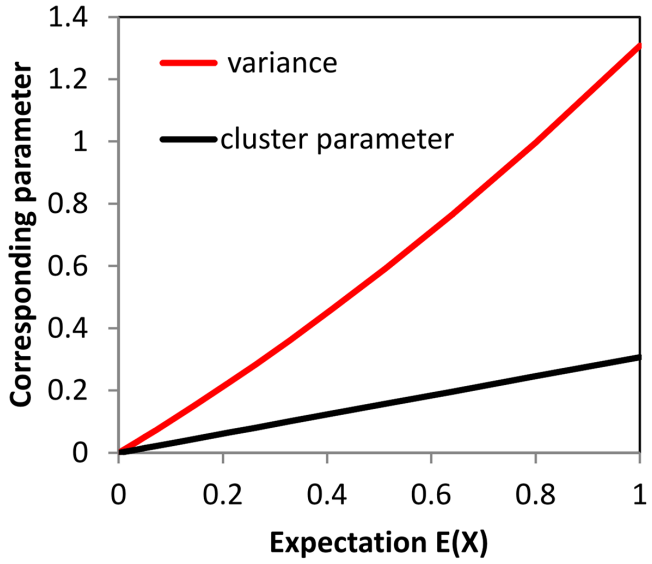

Figure 1. Relation variance and expectation according to Eqs. (10) and (13) with dispersion parameter $\beta = 0.3074$ and corresponding behaviour of cluster parameter ϕ according to Eq. (5).

NHESSD
3, 1775–1787, 2015

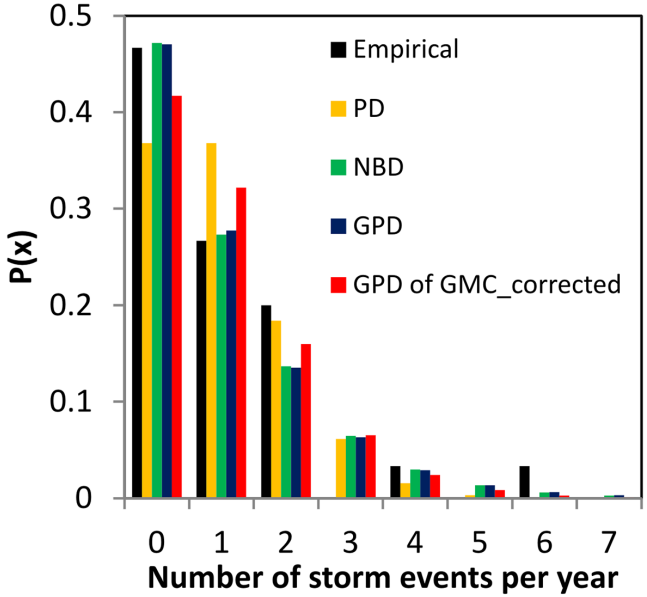

Figure 2. Distributions for the DWD-sample for RL ≥ 1.

NHESSD
3, 1775–1787, 2015

Statistical detection and modeling of the over-dispersion of winter storm occurrence

M. Raschke