Introduction to Stochastic Processes Project Report: modeling Droughts in Brazil using NHPPs

Marcel Braasch
Technical University of Munich
marcelbraasch@gmail.com

Nicola Dal Cin
University of Udine
dalcin.nicola@spes.uniud.it

Marco Hövekenmeier
Technical University of Munich
m.hoevekenmeier@tum.de

1 Introduction

In this report we showcase our results for assessing the work of Achcar et al. (2016), *Use of non-homogeneous Poisson process (NHPP) in presence of change-points to analyze drought periods: a case study in Brazil*.

The project report is structured as following. In Section 2 we explain the subject of this study, namely the SPI data. Section 3 serves as a refresher on the methods used throughout this work. Equipped with the proper tools, Section 4 defines the model as used in the work. In Section 5 we explain our extensions to the work. We introduce the Bayesian framework used to assess the model in Section 6. We present and discuss the results in Section 7 and wrap up the report with a conclusion in Section 8.

Further, we make our code available for possible extensions[1].

2 Data

Droughts and their frequency are a well studied object in literature. Usually, one can motivate the topic from three perspectives, namely, from a meteorologic, agricultural or hydrological point of view (Wang et al., 2021). In our case we regard it as shortage in some component of the hydrological cycle through missing precipitation (Tate and Gustard, 2000).

Achcar et al. (2016) decide to study this behaviour in Campinas, Brazil, where occurrences of droughts are regarded as behaviour deviating from the norm. In other regions in Brazil, droughts are normal incidents. Likely due to global warming, changes of weather behaviour have been observed in recent years. These changes can be seen through the data set at hand.

2.1 Standard Precipitation Index

The Standard Precipitation Index (SPI) is a simple and effective index to analyze wet periods or dry periods, first introduced by McKee et al. (1993). One can compute various time scales of the index leading to different meteorological interpretations. Typically, one distinguishes among $SPI - i$ with $i \in \{1, 3, 6, 12, 24\}$. The indices are calculated such that cumulative probability of rainfall occurring at a climate station is fitted to a gamma distribution, where values for a specific point in time are averaged and normalized over a window i. For a more thorough procedure, please refer to (McKee et al., 1993).

A table with a possible weather classification based on SPI values can be found in Table 1.

SPI	Meaning
$x \geq 2.0$	Extremely wet
$2.0 > x \geq 1.5$	Very wet
$1.5 > x \geq 1.0$	Moderately wet
$1.0 > x \geq -1.0$	Near normal
$-1.0 > x \geq -1.5$	Moderately dry
$-1.5 > x \geq -2.0$	Severely dry
$x < -2.0$	Extremely dry

Table 1: SPI values and their respective meanings.

2.2 Dataset

The dataset for this study are the $SPI-1$, $SPI-3$, $SPI-6$ and $SPI-12$ records for the city Campinas in Brazil, from January 1, 1947 to May 1, 2011. Data can be downloaded from here[2], however, requires post–processing and cleaning. A ready-to-use version is provided in our repository. In Figure 1 one can observe plots of the time series data. In Figure 2 one can see the cumulative counts of droughts, increased by 1 every time the

[1] github.com/marcoHoev/nhpp_drought_modeling

[2] http://www.ciiagro.sp.gov.br/ciiagroonline/Listagens/SPI/LspiLocal.asp

respective SPI value is below -1.0. In other words, we consider a drought event to happen anytime the SPI is below that threshold.

3 Background

In this section we briefly recall some basic concepts in stochastic processeses, especially concerning non-homogeneous Poisson processes and power law processeses.

3.1 Poisson Processes

Poisson processes may have various realizations and definitions, depending on the chosen mathematical rigour and domain of interest. In our case, we regard a process on the positive half-line. This way it can be interpreted as a counting or queuing process (Last and Penrose, 2017). Intuitively, one can regard it as counting process representing the total number of occurrences or events that have happened within a fixed interval $(a, b]$. Formally, this can be expressed as the following.

Definition 1. *A counting process* $\{N(t),\ t \geq 0\}$ *is called a* non-homogeneous Poisson process *(NHPP) with mean value function* $m(t) = \int_0^t \lambda(x)dx$ *if*

$$N(0) = 0, \tag{1}$$

$$\{N(t), t \geq 0\} \text{ has independent increments,} \tag{2}$$

$$P(N(t+h) - N(t) = 1) = \lambda(t)h + o(h), \tag{3}$$

$$P(N(t+h) - N(t) \geq 2) = o(h), \tag{4}$$

where $\lambda(t) > 0$ *is called the intensity function.*

If the density is constant, i.e. $\lambda(t) := \lambda t$, we regard the process a *homogeneous Poisson process* (HPP).

There are various interesting properties to note. First of all, as the name suggests, it is possible to show that the amount of events occurring in a certain interval follow a Poisson distribution. This is formalized by the following result, which is the one used in our implementation.

Theorem 1. *A non-homogeneous Poisson process* $\{N(t),\ t \geq 0\}$ *satisfies*

$$N(t+s) - N(t) \sim Pois(m(t+s) - m(t)) \tag{5}$$

for all $t, s \geq 0$.

Moreover, note that if the underlying process is an HPP, then it is clear that the Poisson distribution of the number of arrivals in each interval $(t+s, t]$ does not depend on t, but on the interval length s (Stoyan et al., 2013; Daley and Vere-Jones, 2008).

3.2 Likelihood function

In Bayesian statistics (which will be our tool of choice), it is common to regard the likelihood function as a means sample information influencing the posterior probability of the parameters under current assumptions (Zellner, 1996). Formally, the likelihood of a NHPP can be expressed as the following.

Definition 2. *The likelihood function for* $\boldsymbol{\theta}$ *of the truncated model at time* T *is*

$$L(\boldsymbol{\theta}|D_T) = \prod_{i=1}^{n} \lambda(t_i) \cdot \exp\left[-m(T)\right], \tag{6}$$

where $D_T = \{n; t_1, \ldots, t_n; T\}$ *is the data set.*

3.3 Power Law Process

In this report, we study the occurrence of droughts. Typically, events of this type are regarded as rare. This suggests the usage of mechanism which are able to model atypical behaviour. A common tool in probabilistic and statistical modeling in this domain is the power law process. Besides its application to drought research (Naumann et al., 2015), it has been frequently applied to related areas such as excessive rainfalls (Salles et al., 2002; Zhang et al., 1999) or volcanic eruptions (Eadie and Favis-Mortlock, 2010). Formally, the power law process can be defined as the following.

Definition 3. *An NHPP* $\{N(t),\ t \geq 0\}$ *is a* power law process *(PLP) with parameter* $\boldsymbol{\theta} := (\alpha, \sigma)$ *if its mean function is*

$$m_{PLP}(t; \boldsymbol{\theta}) = \left(\frac{t}{\sigma}\right)^{\alpha}, \quad \alpha, \sigma > 0, \tag{7}$$

thus it follows that the intesity of the PLP is given as

$$\lambda_{PLP}(t; \boldsymbol{\theta}) = \left(\frac{\alpha}{\sigma}\right)\left(\frac{t}{\sigma}\right)^{\alpha-1}. \tag{8}$$

3.4 Change Points

When analyzing data it may occur that we observe abrupt changes in the behaviour of the observed distribution. Thus models may benefit from a change in behaviour. Change points have been a frequent subject of study in the previous years (Matthews and Farewell, 1982). Especially in Bayesian approaches it is of importance to model the change point as part of the parameters and infer it from the data (Pievatolo and Ruggeri, 2004; Dey and Purkayastha, 1997).

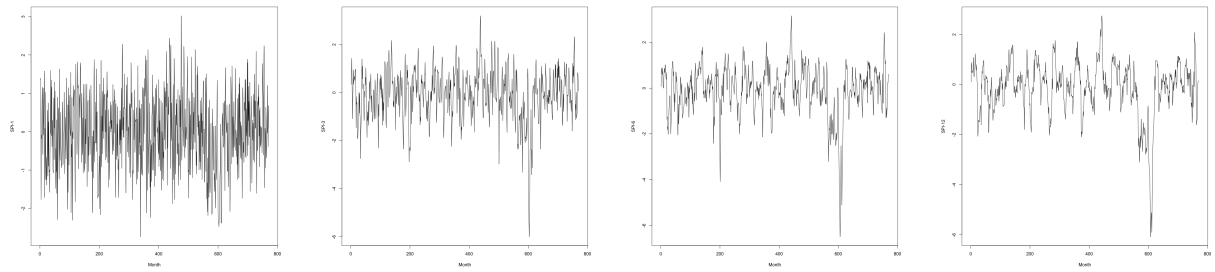

Figure 1: Time series plots of the SPI-1, SPI-3, SPI-6 and SPI-12 data (from left to right).

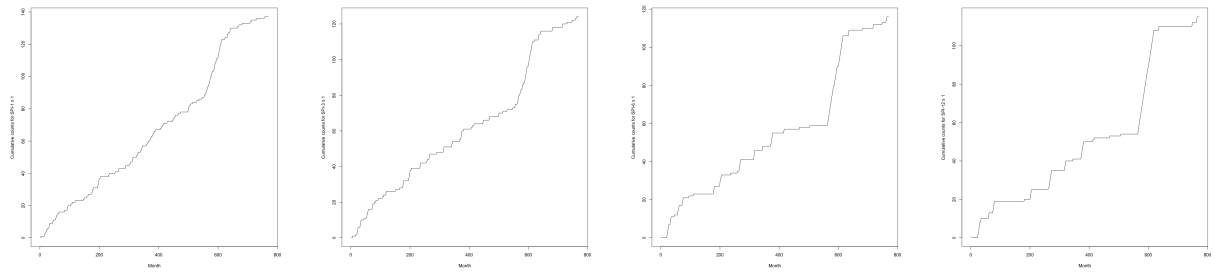

Figure 2: Cumulative plots for theme series plots of the SPI-1, SPI-3, SPI-6 and SPI-12 data (from left to right). The count is increased by 1, every time the respective SPI value is below -1.0.

4 Model

Equipped with the data and general definitions we can move on to define the model used in this work. Achcar et al. (2016) distinguish between three types of models to fit to the data, that is, **(1)** a PLP with no change points, **(2)** a PLP with one change point and **(3)** a PLP with two change points.

We assume that the number of droughts up to a certain month follows a counting process. The data set can be interpreted as a specific realization of this process. We assume that the counting process is (or is approximated by) one of the above cases **(1)** - **(3)**, and perform a Bayesian analysis to determine the parameters that provide a good fit to the data.

In the case **(1)** we simply consider a PLP model as in Def. 3. To express the model with one change point, instead, we need to adapt the mean in the following way.

Definition 4. *The mean* $m_{PLP}(t;\boldsymbol{\theta})$ *for the PLP with one change point* τ *is defined as*

$$\left(\frac{t}{\sigma_1}\right)^{\alpha_1}$$

if $0 \leq t \leq \tau$, *and*

$$\left(\frac{t}{\sigma_1}\right)^{\alpha_1} + \left(\frac{\tau}{\sigma_2}\right)^{\alpha_2} - \left(\frac{t}{\sigma_2}\right)^{\alpha_2} \tag{9}$$

otherwise.

Moreover, to express the model with two change points the adaptation is the following.

Definition 5. *The mean* $m_{PLP}(t;\boldsymbol{\theta})$ *for the PLP with two change points* (τ_1, τ_2) *is defined as*

$$\left(\frac{t}{\sigma_1}\right)^{\alpha_1}$$

if $0 \leq t \leq \tau_1$,

$$\left(\frac{\tau}{\sigma_1}\right)^{\alpha_1} + \left(\frac{t}{\sigma_2}\right)^{\alpha_2} - \left(\frac{\tau_1}{\sigma_2}\right)^{\alpha_2}$$

if $\tau_1 \leq t \leq \tau_2$, *and*

$$\left(\frac{\tau_1}{\sigma_1}\right)^{\alpha_1} + \left(\frac{t}{\sigma_3}\right)^{\alpha_3} - \left(\frac{\tau_2}{\sigma_3}\right)^{\alpha_3} + \left(\frac{\tau_2}{\sigma_2}\right)^{\alpha_2} - \left(\frac{\tau_1}{\sigma_2}\right)^{\alpha_2} \tag{10}$$

otherwise.

Extended definitions for the intensities and likelihoods are neglected for the sake of conciseness and easily derivable from Eq. (9) and (10), otherwise we refer to Achcar et al. (2016).

The models introduced in the paper of Achcar et al. (2016) are fitted for the respective SPI values. Plots of the results can be found in Figure 3.

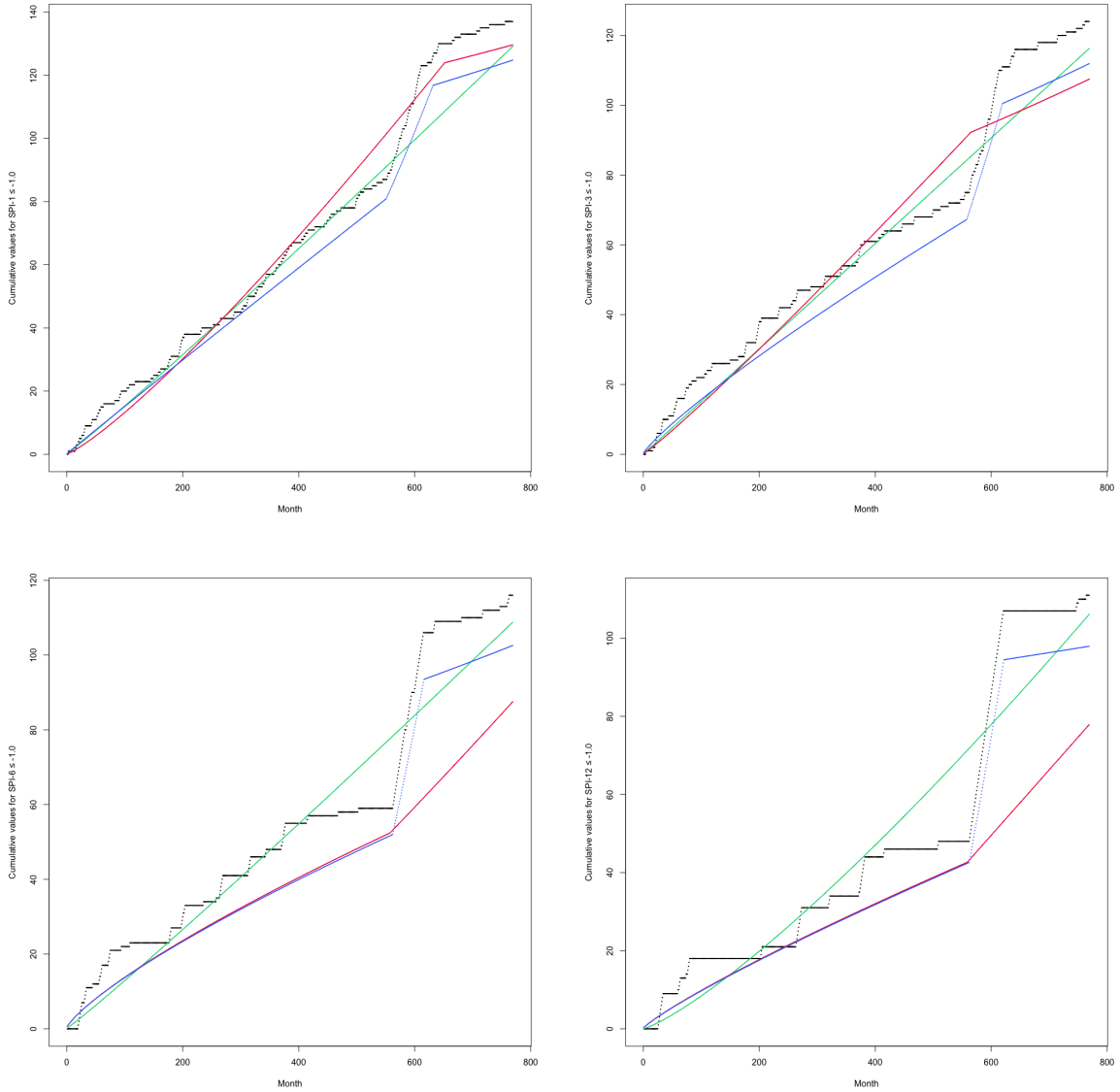

Figure 3: Plots of the cumulative SPI-1 (top left), SPI-3 (top right), SPI-6 (bottom left) and SPI-12 (bottom right) with the mean of the estimated processes. The model with no change point is shown in green, the model with one change point is shown in red and the model with no change point is shown in blue.

5 Extension

We challenge the model defined by Achcar et al. (2016) by a simple, adaptable method based on a polynomial formulation. We notice various trends in the underlying data which we hypothesize to be accurately simulated by polynomials of various degrees. Formally, we express this as the following.

Definition 6. *We call an NHPP* $\{N(t),\ t \geq 0\}$ *an* N*-th* degree polynomial process *(PP) with parameter* $\boldsymbol{\theta} := (\alpha_1, \ldots, \alpha_N)$ *if its mean function is*

$$m_{PP_N}(t; \boldsymbol{\theta}) = \sum_{i=1}^{N} \alpha_i t^i \tag{11}$$

Notice in Figure 3, for SPI-1, top left, that the mean of the model with two change points (blue) is lying slightly under the data plot. We believe that the PLP has issues fitting the data accurately, possibly due to numerical reasons. Thus for SPI-1, we set N=1 (i.e., define an HPP), using two change points. We call this process *linear process*.

For SPI-3, 6 and 12 we identify possible quadratic trends, setting N=2, using two change

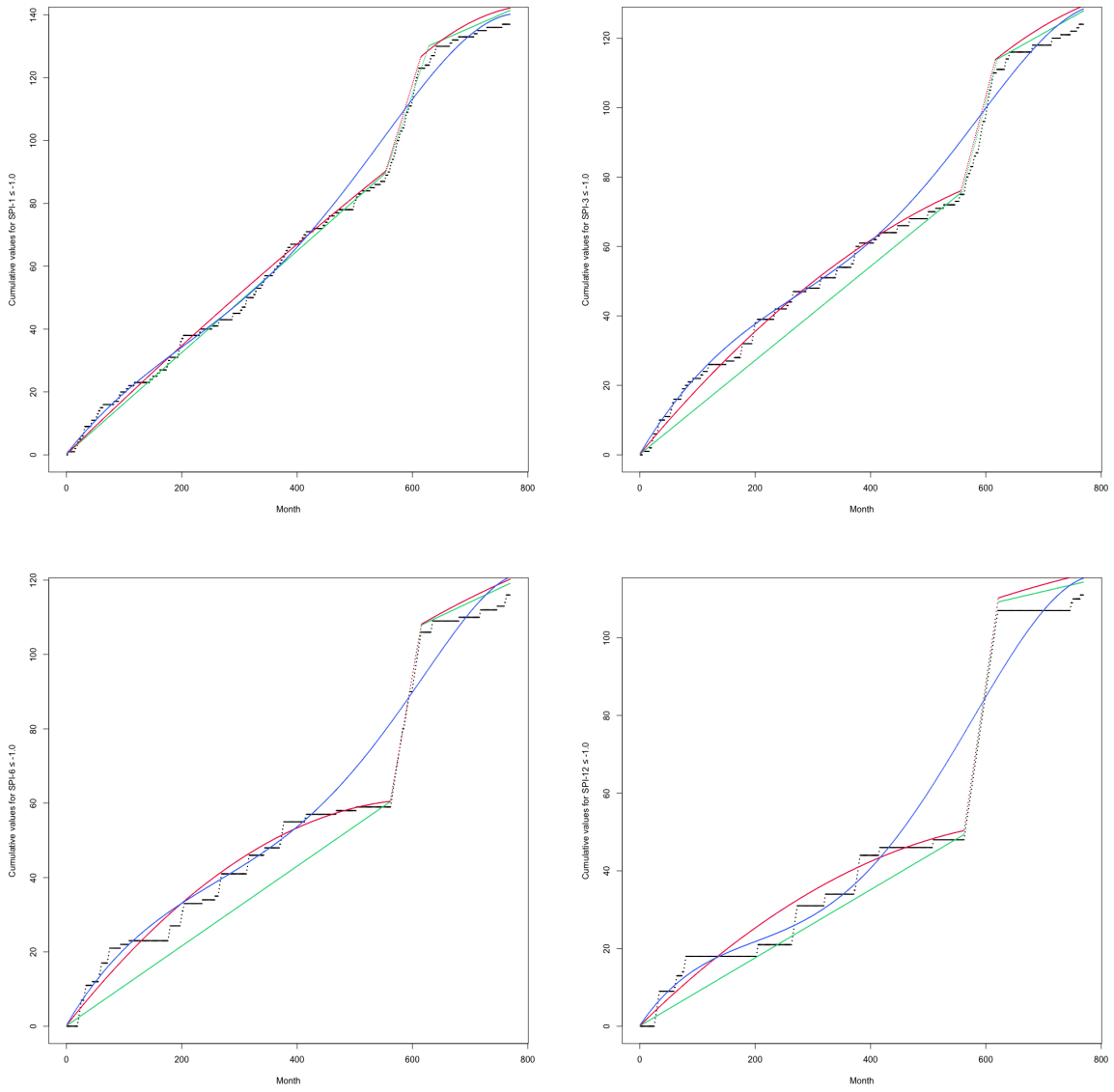

Figure 4: Plots of the cumulative SPI-1 (top left), SPI-3 (top right), SPI-6 (bottom left) and SPI-12 (bottom right) with the mean of the estimated processes. PP_1 with two change points is shown in green, PP_2 with two change points is shown in red and the PP_8 with no change point is shown in blue.

points. We call this process *quadratic process* (QP).

Our last experiments consists of setting $N{=}8$. Our hope is to obtain a polynomial flexible enough to fit the data without having to introduce any change points. We call this process *8-th degree polynomial process*.

For the *quadratic* and *8-th degree polynomial process* we choose eight parameters to match the complexity of the PLP with two change points. The plots of the mean functions for our models can be found in Figure 4.

6 Bayesian Analysis

To perform inference on the parameters, similar to Achcar et al. (2016), we utilize a Bayesian approach. The posterior distribution of the parameters can be expressed as

$$p(\boldsymbol{\theta}|D_t) \propto p(\boldsymbol{\theta})L(\boldsymbol{\theta}|D_t) \tag{12}$$

where $p(\boldsymbol{\theta})$ is the prior belief about the parameters, L is the likelihood as defined in Eq. (2), and D_t is the dataset.

For all models, as the authors, we assume non-informative independent priors. For the PLP without change points we assume α, $\sigma \sim U[0, 100]$.

For the PLP with one change point we assume α_1, α_2, σ_1, $\sigma_2 \sim U[0, 100]$ and $\tau \sim U[0, T]$. For the PLP with two change points we assume α_1, α_2, α_3, σ_1, σ_2, $\sigma_3 \sim U[0, 100]$. For PP_1 with two change points we assume $\alpha_{1,1}$, $\alpha_{1,2}$, $\alpha_{1,3} \sim U[0, 100]$ with $\alpha_{i,j}$ following the definition of the mean in Eq. (11) before the j-th change point. For PP_2 with two change points we assume $\alpha_{1,1}$, $\alpha_{1,2}$, $\alpha_{1,3} \sim U[0, 100]$ and $\alpha_{2,1}$, $\alpha_{2,2}$, $\alpha_{2,3} \sim U[-1, 1]$ for better convergence. Note that since the optimal values found for these parameters are very close to 0, this can still be considered a non-informative prior. For PP_8 with no change points we assume α_1, α_2, α_3, α_4, α_5, α_6, α_7, $\alpha_8 \sim U[-1000, 1000]$. For all models with two change points we assume $\tau_1 \sim U[400, 600]$ and $\tau_2 \sim U[600, T]$ as suggested in the article cited.

According to this formulation, parameters are estimated using Markov Chain Monte Carlo. We decide to implement the sampling procedure using `r2jags`[3].

7 Results

In this section we compare the models by the means of the Deviance Information Criterion (DIC) (Spiegelhalter et al., 2002). Note that smaller values of DIC indicate better performances. Results can be viewed in Table 2. The top table shows results for the models defined by Achcar et al. (2016), the bottom table shows results for the polynomial processes defined in Eq. (11).

SPI	$\text{PLP}_{2,0}$	$\text{PLP}_{4,1}$	$\text{PLP}_{6,2}$
1	751.3	739.5	717.6
3	705.3	705.1	660.1
6	675.4	655.9	572.9
12	652.5	620.9	499.8

SPI	$\text{LP}_{3,2}$	$\text{QP}_{6,2}$	$\text{PP}_{8,0}$
1	714.8	**714.3**	736.5
3	662.2	**657.3**	696.2
6	577.6	**564.1**	663.7
12	500.8	**498.8**	627.5

Table 2: DIC comparison between the PLP, the PP with $N = 1$ (LP), the PP with $N = 2$ (QP) and the PP with $N = 8$ (PP). $\text{P}_{a,b}$ denotes having a parameters with b change points.

Noticeably, one can see that change points increase model performances significantly within the class of PLP models. Further, $\text{PLP}_{6,2}$ provides better results than $\text{PP}_{8,0}$ and $\text{LP}3, 2$. Our hypothesis of a higher order polynomial mean without change points being flexible enough to fit the data cannot be validated. The poor fit of the model can be seen in Figure 4, especially after $t \geq 400$ where other models correctly estimate change points.

For the case of drought modeling, using the PLP does not appear to be the single optimal choice. The $\text{QP}_{6,2}$ method results in better DIC values for all SPI scales considered. Especially for SPI-6 and SPI-12, one can see how well the QP fits the data in the range up to the first change point.

Furthermore, it is worth noting how close the DIC values of the HPP are compared to the PLP.

For the exact values of the parameters for each polynomial process please refer to Appendix A, Tables 1, 2 and 3. The values of the parameters for the PLP can be found in Achcar et al. (2016).

8 Conclusion

In this work, we have evaluated and resumed the work of Achcar et al. (2016). We model processes of counting droughts according to data based on SPI. The models compared in this report are power law processes, with and without change points (Achcar et al. (2016)'s models), and polynomial processes (our extensions). We carry out extensive experiments, compare results and show a variety of interesting insights. Assuming the presence of change points is very useful and can significantly improve the estimation of the underlying process. Moreover, we report that NHPPs using polynomial means yield even better models to fit the data.

Generalizing this type of analysis possibly requires a variable number of change points. Achcar et al. (2016) explicitly encode their assumed knowledge of change points into the prior distributions. Changing behaviour in the data is highly dependent on the local features of the geographic area in exam. In the best case, one would automatically infer them from the data. Finding change points flexibly seems like a challenging extension.

A further extension of this work could be the application of the proposed models to other data sets on a large scale. This enables a comparison of our results with the work of Achcar et al. (2016).

[3] https://cran.r-project.org/web/packages/R2jags/

References

Jorge Alberto Achcar, Emílio Augusto Coelho-Barros, and Roberto Molina de Souza. 2016. Use of non-homogeneous poisson process (nhpp) in presence of change-points to analyze drought periods: a case study in brazil. *Environmental and Ecological Statistics*, 23(3):405–419.

Daryl J Daley and David Vere-Jones. 2008. *An Introduction to the Theory of Point Processes. Volume II: General Theory and Structure*. Springer.

Dipak K Dey and Sumitra Purkayastha. 1997. Bayesian approach to change point problems. *Communications in Statistics-Theory and Methods*, 26(8):2035–2047.

Chris Eadie and David Favis-Mortlock. 2010. Estimation of drought and flood recurrence interval from historical discharge data: a case study utilising the power law distribution. In *EGU General Assembly Conference Abstracts*, page 13568.

Günter Last and Mathew Penrose. 2017. *Lectures on the Poisson process*, volume 7. Cambridge University Press.

David E Matthews and Vernon T Farewell. 1982. On testing for a constant hazard against a change-point alternative. *Biometrics*, pages 463–468.

Thomas B McKee, Nolan J Doesken, John Kleist, et al. 1993. The relationship of drought frequency and duration to time scales. In *Proceedings of the 8th Conference on Applied Climatology*, volume 17, pages 179–183. Boston, MA, USA.

Gustavo Naumann, Jonathan Spinoni, Jürgen V Vogt, and Paulo Barbosa. 2015. Assessment of drought damages and their uncertainties in europe. *Environmental Research Letters*, 10(12):124013.

Antonio Pievatolo and Fabrizio Ruggeri. 2004. Bayesian reliability analysis of complex repairable systems. *Applied Stochastic Models in Business and Industry*, 20(3):253–264.

Christian Salles, Jean Poesen, and Daniel Sempere-Torres. 2002. Kinetic energy of rain and its functional relationship with intensity. *Journal of Hydrology*, 257(1-4):256–270.

David J Spiegelhalter, Nicola G Best, Bradley P Carlin, and Angelika Van Der Linde. 2002. Bayesian measures of model complexity and fit. *Journal of the royal statistical society: Series b (statistical methodology)*, 64(4):583–639.

Dietrich Stoyan, Wilfrid S Kendall, Sung Nok Chiu, and Joseph Mecke. 2013. *Stochastic geometry and its applications*. John Wiley & Sons.

EL Tate and Alan Gustard. 2000. Drought definition: a hydrological perspective. In *Drought and drought mitigation in Europe*, pages 23–48. Springer.

Qianfeng Wang, Jingyu Zeng, Junyu Qi, Xuesong Zhang, Yue Zeng, Wei Shui, Zhanghua Xu, Rongrong Zhang, Xiaoping Wu, and Jiang Cong. 2021. A multi-scale daily spei dataset for drought characterization at observation stations over mainland china from 1961 to 2018. *Earth System Science Data*, 13(2):331–341.

Arnold Zellner. 1996. Introduction to bayesian inference in econometrics.

Wei Zhang, Nader Moayeri, et al. 1999. Power-law parameters of rain specific attenuation.