On using nonparametric approaches for precipitation estimation

Mariusz Grządziel[1,*]

[1]Wrocław University of Environmental and Life Sciences, Department of Mathematics, Grunwaldzka 53, 50357 Wrocław, Poland

Abstract. Nonparametric density estimation methods have been used for precipitation estimation for decades. The new approach for estimating the density proposed recently by Geenens and Wang appears to offer advantages over them as far as their behavior in the tail is concerned. Real data analyses presented in this paper confirm the usefulness of this new approach for precipitation modelling.

1 Introduction

Kernel density estimation has many applications in environmental sciences; compare e.g. [1–3]. However, in many practical situations, e.g. when we want to estimate precipitation distribution, we have to estimate density of positive variables which may be challenging. It is known that the kernel estimators have heavy bias at and near the boundary; compare [4, page 594]. Some remedies for this defect and other drawbacks of kernel estimators that arise in the case when the random variable of interest is positive were proposed in the literature. However, these remedies are not quite satisfactory and the problem of density estimation of positive random variables is still an area of active research.

We will present some new methods addressing this problem in Section 2; among them the method proposed by Geenens and Wang [5] deserves in our opinion particular interest. Its usefulness will be illustrated on real datasets in Section 3. Conclusions will be given in the final section.

2 Density estimation of positive variables

For a given a sample $X_1, \dots, X_n$ consisting of i.i.d. copies of a random variable X with unknown distribution P_X admitting a density f_X, the kernel of estimator $\hat{f}_X$ takes the form

$$\hat{f}_X(x) = \frac{1}{nh}\sum_{k=1}^{n} K\left(\frac{x - X_k}{h}\right), \tag{1}$$

where the kernel K is a non-negative function satisfying the conditions

$$\int_{-\infty}^{\infty} K(x)dx = 1, \tag{2}$$

$$\int_{-\infty}^{\infty} x\,K(x)dx = 0 \tag{3}$$

and $h > 0$ is known as the bandwidth. K is usually assumed to be symmetric about 0.

The statistical properties of the estimator (1) have been presented e.g. in [6]. It has been shown that it reliably estimates f_X when it is supported on $\mathbb{R}$. If the support of f_X is bounded, the kernel estimator $\hat{f}_X$ may have some drawbacks.

In this section we consider the case when P_X is the distribution of a positive continuous random variable X having density f_X supported on $\mathbb{R}^+ = (0, +\infty)$. Such variables naturally arise in various areas of environmental sciences.

As it already has been mentioned, the kernel estimator (1) usually fails to correctly estimate the behaviour of f_X near 0; also its tail behaviour often is not satisfactory. In the neighbourhood of 0 the estimator may place positive probability mass on the negative half-line which results in the *boundary bias*. In the tail region the estimator may produce artificial local maxima at each observation, sometimes referred to as *spurious bumps*; compare [7]. Several remedies to overcome the boundary bias, such as the reflection method [8], were proposed in 1970-ies and 1980-ies.

In 2000 Chen [9] addressed both mentioned problems arising in the case when f_X is supported on $\mathbb{R}^+ = (0, +\infty)$ by introducing the asymmetric kernel density estimator of f_X

$$\hat{f}_X(x) = \frac{1}{n}\sum_{k=1}^{n} L_{x,h}(X_k), \tag{4}$$

where $L_{x,h}$ is an asymmetric $\mathbb{R}^+$-supported density whose parameters are functions of x and h. He proposed two 'asymmetric kernels':

[*] Corresponding author: mariusz.grzadziel@upwr.edu.pl

$$L^{1}_{x,h}(t) = g_{x/h+1,h}(t), \tag{5}$$
$$L^{2}_{x,h}(t) = g_{\rho(x,h),h}(t), \tag{6}$$

where $g_{\alpha,b}$ is the density of the Gamma distribution with the shape parameter α and the scale parameter b while

$$\rho(x,h) = \begin{cases} x/h & \text{if } x \geq 2h; \\ \frac{1}{4}(x/h)^2 + 1 & \text{if } x \in [0,2h). \end{cases} \tag{7}$$

Several types of asymmetric kernels were investigated in the subsequent literature [10–13]. However, it seems that the estimators using these kernels do not show significant advantage over the Chen estimators.

Another approach to estimating f_X supported on $\mathbb{R}^+ = (0, +\infty)$ consists in estimating the density of $Y = \log X$ and obtaining an estimator of the density of X using well-known results concerning functions of random variables.

The fact that

$$f_X(x) = \frac{f_Y(\log\ x)}{x} \tag{8}$$

for all $x > 0$ suggests an estimator of f_X having the form

$$\hat{f}_X(x) = \frac{\hat{f}_Y(\log\ x)}{x} \quad \text{for } x > 0. \tag{9}$$

Here $\hat{f}_Y$ is a suitably chosen estimator of the density f_Y of the random variable Y supported on $\mathbb{R}$. We will refer to the estimator (9) as the log-transform estimator of the density function f_X; compare [14].

Geenens and Wang [5] pointed out that choosing a kernel estimator for estimating f_Y may result in some drawbacks. This is due to the fact that the kernel estimators are known to poorly estimate tails of densities, compare [4, page 594]. One way to overcome this shortcoming is to use the local likelihood density estimator of f_Y (instead of the kernel estimator). The estimator constructed using this approach has many favourable properties; compare [5]. A simulation study in [5, Section 7] indicates that when the degree of the polynomial that approximates the logarithm of the density of Y is 2 (compare [5, Section 2.3]), this estimator has small Mean Integrated Absolute Relative Error (MIARE); for an estimator $\hat{f}_X$ of the density f_X of a positive random variable X MIARE is defined as

$$\mathbb{E}\left(\int_0^\infty \frac{|\hat{f}_X(x) - f_X(x)|}{f_X(x)}\,dx\right). \tag{10}$$

This shows that this local log-quadratic transformation kernel density estimator based on the log-transformation (in short: LL-LT estimator) perform well in the tail.

This property was confirmed by the simulation study presented in [15], the version of the preprint [5] that was accepted for publication; in this simulation study other criterion than MIARE, the mean integrated squared error, was taken into account.

3 Applications to precipitation estimation

The applications of nonparametric methods to precipitation estimation were considered in [16]. The authors suggest using the log-transform density estimator (9) in which $\hat{f}_Y$ was computed using the kernel method with bandwidth proposed by Sheather and Jones [17]. The log-transform kernel density estimator was also used in [18] for estimating precipitation amounts for various time intervals.

Using nonparametric methods for estimating precipitation extremes appears to be particularly challenging. The estimator (9) in which $\hat{f}_Y$ is estimated using spline methods seems to be particularly promising for this aim [19].

3.1 Real data analyses

We have performed analyses of three datasets that were labelled according to the places they were collected from:

- The dataset ‘Łódź’ contains the measurements of monthly precipitation in June (in mm) in the years 1954–1992 from the meteorological station Łódź-Lublinek.
- The dataset ‘Wichita’ contains the measurements of monthly precipitation in August (in mm) in Wichita, Kansas, in the years 1980–2011. This dataset is a part of the R package SPEI [20].
- The dataset ‘Kraków’ contains the annual (from May to September) maximal precipitation intensity measurements (mm/h) from Jagiellonian University Astronomical Observatory in Kraków in the years 1961–1985. The intensities are computed for the time interval of 24 hours. The dataset is taken from the monograph [21, page 354].

For the dataset ‘Łódź’ the Gamma distribution and the Lognormal distribution were fitted using the procedure *fitdist* from the package *fitdistrplus* [22]. The LL-LT estimator was computed using the R script accompanying the paper [5] (available at the *Journal of the Comuptational and Graphical Statistics* website).

Table 1. Precipitation (mm) in June in Łódź (1954–1992): quantiles computed for the obtained distributions (estimates of the density function): Lognormal, Gamma, LL-LT, Gamma kernel, Gaussian kernel (with Sheather-Jones bandwidth).

Distribution	Quantile			
	$\Phi(-3)$	0.01	0.99	$\Phi(3)$
Lognormal	13.04	18.49	205.41	291.11
Gamma	8.22	14.51	175.87	221.93
LL-LT	10.49	16.85	193.39	264.47
Gamma kernel	0.75	5.22	226.52	277.29
Gaussian Kernel (SJ)	-3.31	7.36	221.74	236.42

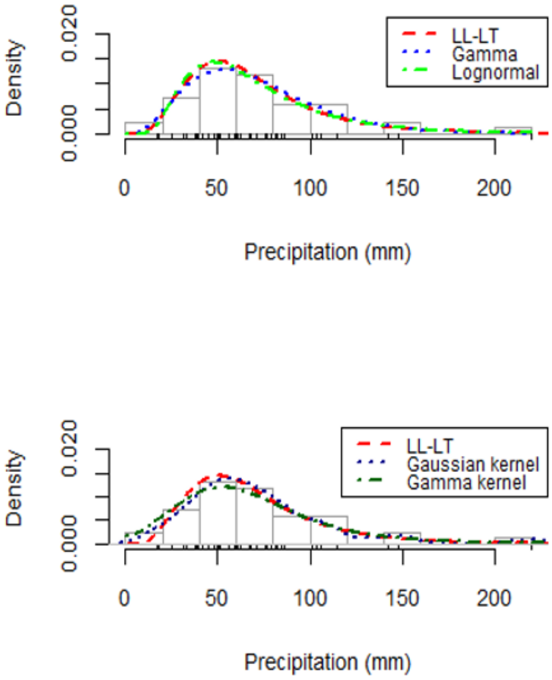

Fig. 1 Dataset 'Łódź': density function estimated using parametric methods (Lognormal, Gamma) and nonparametric methods (LL-LT, Gaussian kernel with Sheather-Jones bandwidth, Gamma kernel).

The estimates are displayed in the Figure 1 (upper panel); the graph of the LL-LT was also displayed in the lower panel of the Figure 1 together with the graphs of the Gamma kernel estimate as well as the Gaussian kernel estimate of the density of the precipitation. The Gamma kernel estimate was computed using the bandwidth determined by the *unbiased cross validation* method implemented in the package *kdensity* [23]. The Gaussian kernel density estimate was computed using the bandwidth determined by the Sheather-Jones method implemented in the bw.SJ procedure from the *R* package *stats* [24]. Some quantiles for the estimated densities were displayed in Table 1; the $\Phi(3)$ and $\Phi(-3)$ quantiles are of interest when computing the standardized precipitation index (SPI) which could be defined as $\Phi^{-1}(F(t))$, where t is the amount of precipitation, Φ denotes the cumulative distribution function of the standard normal distribution and F is the cumulative distribution function of the suitably chosen distribution fitted to data; compare e.g. [25].

We see that the LL-LT estimate did approximately as well as the Log-normal estimate. Let us note that the dataset 'Łódź' was also analyzed in [25]: the authors have demonstrated that the fit obtained for the Log-normal distribution was better than for the competing distributions considered in [25].

Similar results were obtained for the dataset 'Wichita'. The LL-LT and the Gamma estimates appear to perform best; the $\Phi(3)$ quantile for the Log-normal distribution in the Table 2 seems to be unreasonably large. The Gamma kernel estimate did well in the right tail, but its boundary behaviour seems to be less satisfactory; the values of the density function near the boundary seem to be overestimated using this approach.

Table 2. Precipitation (mm) in August in Wichita (1980-2011): quantiles computed for the estimates of the density.

Distribution	Quantile			
	$\Phi(-3)$	0.01	0.99	$\Phi(3)$
Lognormal	5.68	10.03	511.29	903.34
Gamma	2.41	6.84	314.05	421.50
LL-LT	3.21	7.24	342.34	413.48
Gamma ker.	0.28	2.10	331.66	408.31
Gaussian kernel (SJ)	-52.79	-27.52	317.52	353.86

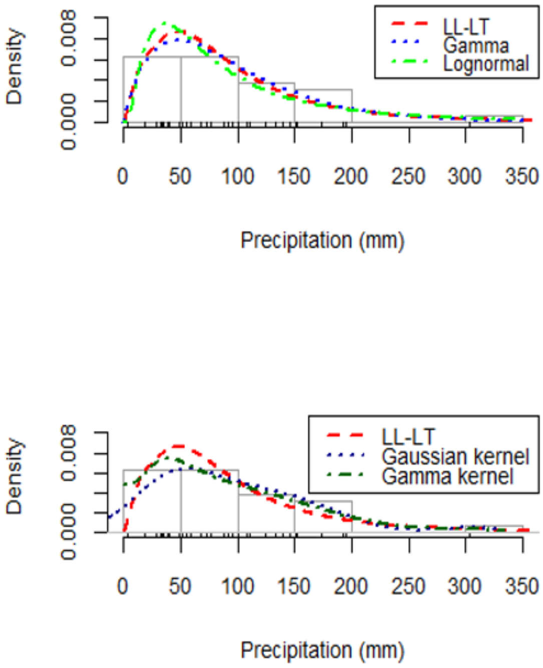

Fig. 2. Dataset 'Wichita' : density function estimated using parametric methods (Lognormal, Gamma) and nonparametric methods (LL-LT, Gaussian kernel with Sheather-Jones bandwidth, Gamma kernel).

Table 3. Maximum rainfall intensity (mm/h) in Kraków (1961–1985) computed for the time interval of 24 hours: quantiles computed for the density estimates.

Distribution	Quantile	
	0.99	0.995
Lognormal	3.39	3.72
Gamma	3.23	3.47
GEV	5.02	6.30
LL-LT	4.21	4.83
Gamma kernel	3.65	3.77
Gaussian kernel (SJ)	3.57	3.93

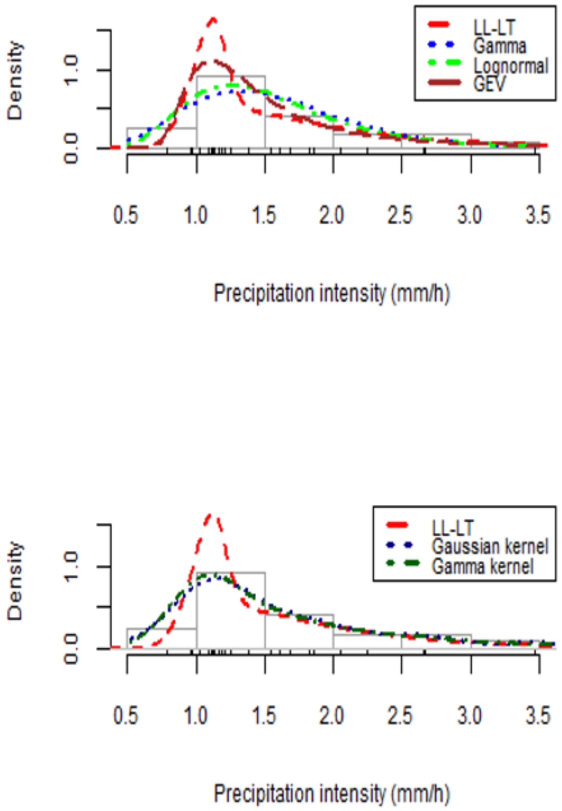

Fig. 3 Dataset 'Kraków': density function estimated using parametric methods (Lognormal, Gamma, GEV) and nonparametric methods (LL-LT, Gaussian kernel with Sheather-Jones bandwidth, Gamma kernel).

The dataset 'Kraków' contains data concerning precipitation extremes. We estimated the density function of the maximal precipitation intensity (mm/h) in the years 1961–1985 with the time interval equal to 24 hours using the same methods as before. Additionally we fit the parameters of the GEV distribution to the data using the maximum likelihood method implemented in the package *evd* [26]. The densities estimates are displayed in Figure 3, the 0.99 and 0.995 quantiles corresponding to the obtained density estimates are presented in Table 3. We see that that the LL-LT and the GEV estimates of the density appear to perform best. It can be thus said that the 'local-likelihood log-transformation approach' proved to be successful also in this case.

4 Conclusions

The real data analyses presented in the previous section confirm that the nonparametric methods may prove to be efficient for precipitation estimation. A new nonparametric approach introduced by Geenens and Wang that bases on data transformation and local likelihood estimation appears to be particularly useful for this aim. There is a need of an extensive simulation study comparing this method with other new promising nonparametric approaches that may be used for precipitation estimation.

References

1. K.D. Kim, J.H. Heo, J. Hydrology **260**, 176–193 (2002)
2. R.A. Ferreyra, G. P. Podestá, C. D. Messina, D. Letson, J. Dardanelli, E. Guevara, S. Meira, Agric. For. Meteorol. **107**, 177–192 (2001)
3. A. Michalski, Meteorol. Hydrol. Water Manage. **4**, 41–46 (2016)
4. S.J. Sheather, Stat. Sci. **19**, 588–597 (2004)
5. G. Geenens, C. Wang, J. Comput. Graph. Stat. (to be published)
6. M.P. Wand, M.C. Jones, *Kernel Smoothing* (Chapman and Hall, Boca Raton, 1995)
7. P. Hall, C. Minnotte, C. Zhang, Ann. Statist. **32**, 2124–2141 (2004)
8. E. Schuster, Comm. Statist. Theory Methods **14**, 1123–1136 (1985)
9. S.X. Chen, Ann. Inst. Statist. Math. **52**, 471–480 (2000)
10. Jin, X. and Kawczak, J. Ann. Econ. Finance **4**, 103–124 (2003)
11. O. Scaillet, J. Nonparametr. Stat. **16**, 217–226 (2004)
12. M. Hirukawa, M. Sakudo, J. Nonparametr. Stat. **27**, 41–63 (2015)
13. G. Igarashi, Commun. Stat. Theory Methods **45**, 6670 – 6687 (2016)
14. A. Charpentier, E. Flachaire, L'Actualité Économique **91**, 141–159 (2015)
15. G. Geenens, C. Wang, *Local-likelihood transformation kernel density estimation for positive random variables* (arXiv preprint, arXiv:1602.04862 v1 [stat.ME], 2016)
16. B. Rajagoplan, U. Lall, D. Tarboton, Stoch. Hydrol. Hydraul. **11**, 523– 547 (1997)
17. S.J. Sheather, M.C. Jones, J.R. Stat. Soc. Ser. B Stat. Methodol. **53**, 683–690 (1991)
18. T. Mosthaf, A. Bardossy, Hydrol. Earth Syst. Sci. **21**, 2643–2481 (2017)
19. W. Huang, D. Nychka, H. Zhang, *Modeling Precipitation Extremes using Log-Histospline* (arXiv preprint, arXiv:1802.09387v1, 2018)
20. S. Beguería, S. Vicente-Serrano, *SPEI: Calculation of the Standardised Precipitation-Evapotranspiration Index* (R package, version 1.7, http://CRAN.R-project.org/package=SPEI, 2017)
21. S. Węglarczyk, *Statistics in environmental engineering* (in Polish) (Technical University in Kraków Press, Kraków, 2010)

22. M.L. Delignette-Muller, C. Dutang, J. Stat. Softw. **64**, 1–34 (2015)
23. J. Moss, M. Tveten, *kdensity: Kernel Density Estimation with Parametric Starts and Asymmetric Kernels* (R package, version 1.0, https://CRAN.R-project.org/package=kdensity, 2018)
24. R Core Team, *R: A language and environment for statistical computing* (R Foundation for Statistical Computing, Vienna, 2015)
25. E. Gąsiorek, E. Musiał, J. Ecol. Eng. **16**, 44–53 (2015)
26. A. Stephenson, R News **2**, 31–32 (2002)