

AN APPLICATION OF THE LIKELIHOOD METHOD TO CHANGE-POINT DETECTION

EDIT GOMBAY¹ AND LAJOS HORVÁTH^{2*}

¹*Department of Mathematical Sciences, University of Alberta, Edmonton, Alberta, Canada T6G 2G1*

²*Department of Mathematics, University of Utah, Salt Lake City, UT 84112, USA*

SUMMARY

We reanalyse the water discharges from the creek Načetínský. We demonstrate that the likelihood method can be used to detect possible changes in the parameters of the distributions of the observations. © 1997 John Wiley & Sons, Ltd.

Environmetrics, **8**, 459–467 (1997)

No. of Figures: 6. No. of Tables: 5. No. of References: 7.

KEY WORDS water discharge; change-point; likelihood function

1. INTRODUCTION

Jarušková (1994) analysed the monthly averages of water discharges from Načetínský measured in l/s during 1951–1990. Načetínský is a small creek in the German part of the Erzgebirge mountains. The forest in the Erzgebirge mountains was heavily damaged by acid rain and it was expected that the large deforestation may have changed the water discharges from Načetínský. Jarušková (1994) assumed that the monthly averages follow log-normal distributions with different means and variances, so using the logarithmic transformation the data were transformed into normal observations. Jarušková (1994) assumed that the transformed series is an autoregressive sequence with no changes at the end of the sequence. She found a change in the mean of the transformed variables and her estimator for the time of change was 1965. She could not detect any changes in the variance of the transformed variables, so there was no change in the shape factor of the original sequence.

In this paper we reanalyse the log-transformed data using likelihood based methods.

2. MONTHS

Let X_1, X_2, \dots, X_n be independent, normal random variables with $EX_i = \mu_i$ and $\text{var } X_i = \sigma^2$. First we wish to test

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_n \text{ and } \sigma_1 = \sigma_2 = \dots = \sigma_n$$

* Correspondence to: L. Horváth, Department of Mathematics, University of Utah, Salt Lake City, UT 84112, USA

Contract grant sponsor: NATO Scientific and Environmental Division

Contract grant number: CRG 960503

CCC 1180–4009/97/050459–09\$17.50

© 1997 John Wiley & Sons, Ltd.

Received 20 August 1996

Revised 28 February 1997

against

$$H_1^{(1)} : \sigma_1 = \sigma_2 = \dots = \sigma_n \text{ and there is an integer } k^* \text{ such that} \\ \mu_1 = \mu_2 = \dots = \mu_{k^*} \neq \mu_{k^*+1} = \dots = \mu_n.$$

This means that the mean changed under $H_1^{(1)}$ while the variance remained constant. Assuming that $k = k^*$, let $\Lambda_k^{(1)}$ denote the likelihood ratio of constant mean against the change in the mean after X_k (the variance is a constant nuisance parameter). Yao and David (1984) showed that under H_0

$$\lim_{n \rightarrow \infty} P \left\{ (2 \log \log n)^{1/2} T_n^{(1)} \leq x + 2 \log \log n + \frac{1}{2} \log \log \log n - \frac{1}{2} \log \pi \right\} = \exp(-2e^{-x}) \quad (1)$$

for all x , where

$$T_n^{(1)} = \max_{1 \leq k < n} (-2 \log \Lambda_k^{(1)})^{1/2}.$$

Let $\{B(t), 0 \leq t \leq 1\}$ be a Brownian bridge and for any $0 < \alpha < 1$ define

$$u(h, l; \alpha) = \sup \left\{ x : P \left\{ \sup_{h \leq t \leq 1-l} \frac{|B(t)|}{(t(1-t))^{1/2}} \leq x \right\} = 1 - \alpha \right\}.$$

Gombay and Horváth (1996a) proved that under H_0 we have

$$\lim_{n \rightarrow \infty} P\{T_n^{(1)} > u(h, l; \alpha)\} = \alpha, \quad (2)$$

if $h = l = (\log n)^\gamma/n$ with some $\gamma > 0$. For the computation of $u(h, l; \alpha)$ we refer to Vostrikova (1981), Gombay and Horváth (1996a) and Csörgö and Horváth (1997).

Next we consider H_0 against

$$H_1^{(2)} : \mu_1 = \mu_2 = \dots = \mu_n \text{ and there is an integer } k^* \text{ such that} \\ \sigma_1 = \sigma_2 = \dots = \sigma_{k^*} \neq \sigma_{k^*+1} = \dots = \sigma_n.$$

Assuming again that $k^* = k$ is known, we compute $\Lambda_k^{(2)}$, the likelihood of constant variance versus two different variances after the k th observation. Gombay and Horváth (1996a) proved that

$$\lim_{n \rightarrow \infty} P \left\{ (2 \log \log n)^{1/2} T_n^{(2)} \leq x + (2 \log \log n)^{1/2} + \frac{1}{2} \log \log \log n - \frac{1}{2} \log \pi \right\} = \exp(-2e^{-x}) \quad (3)$$

and

$$\lim_{n \rightarrow \infty} P\{T_n^{(2)} > u(h, l; \alpha)\} = \alpha \quad (4)$$

if H_0 holds and

$$T_n^{(2)} = \max_{1 \leq k < n} (-2 \log \Lambda_k^{(2)})^{1/2}.$$

Our last alternative is

$$H_1^{(3)} : \text{there is an integer } k^* \text{ such that } \mu_1 = \mu_2 = \dots = \mu_{k^*} \neq \mu_{k^*+1} = \dots = \mu_n \\ \text{and/or } \sigma_1 = \sigma_2 = \dots = \sigma_{k^*} \neq \sigma_{k^*+1} = \dots = \sigma_n.$$

Under $H_1^{(3)}$ at least one of the parameters changed. If the time of change, $k = k^*$ is known then we can compute the likelihood ratio $\Lambda_k^{(3)}$. Let $\{B_1(t), 0 \leq t \leq 1\}$ and $\{B_2(t), 0 \leq t \leq 1\}$ be two independent Brownian bridges and define

$$v(h, l; \alpha) = \sup \left\{ x : P \left\{ \sup_{k \leq t \leq 1-l} \left(\frac{B_1^2(t) + B_2^2(t)}{t(1-t)} \right)^{1/2} \leq x \right\} = 1 - \alpha \right\}$$

for any $0 < \alpha < 1$. It follows from Gombay and Horváth (1996a) that

$$\lim_{n \rightarrow \infty} P[(2 \log \log n)^{1/2} T_n^{(3)} \leq x + 2 \log \log n + \log \log \log n] = \exp(-2e^{-x}) \tag{5}$$

and

$$\lim_{n \rightarrow \infty} P[T_n^{(3)} > v(h, l; \alpha)] = \alpha \tag{6}$$

if $h = l = (\log n)^{\gamma/n}$ with some $\gamma > 0$, where

$$T_n^{(3)} = \max_{1 \leq k < n} (-2 \log \Lambda_k^{(3)})^{1/2}.$$

Vostrikova (1981), Gombay and Horváth (1996a) and Csörgö and Horváth (1997) contain some useful formulas for the computation of $v(h, l; \alpha)$.

We apply this model to the water discharges in the months of January, February, ... and December separately. We have $n = 40$ observations in each case. Table I contains selected critical values for $T_{40}^{(1)}$ and $T_{40}^{(2)}$ when $\alpha = 0.1, 0.05$ and 0.01 . We computed $u(h, l; \alpha)$ with $k = l = (\log n)^{3/2}/n$ (u^*) and $h = l = (\log n)/n$ (u^{**}) when $n = 40$. We used (1) and (3) to get the asymptotical critical values (Asymp. in Table I). $z_{40}^{(1)}(\alpha)$ and $z_{40}^{(2)}(\alpha)$ are simulated critical values for $T_{40}^{(1)}$ and $T_{40}^{(2)}$. Since the distributions of $T_{40}^{(1)}$ and $T_{40}^{(2)}$ do not depend on the values of the mean and the variance under H_0 , we used standard normal random variables. The simulations were repeated 10,000 times. Table II was obtained in a similar fashion using (5) and (6).

Table I. Selected critical values for $T_{40}^{(1)}$ and $T_{40}^{(2)}$

α	u^*	u^{**}	Asymp.	$z_{40}^{(1)}(\alpha)$	$z_{40}^{(2)}(\alpha)$
0.1	2.65	2.80	3.17	2.78	2.83
0.05	2.94	3.07	3.61	3.07	3.18
0.01	3.49	3.60	4.62	3.70	3.96

Table II. Selected critical value for $T_{40}^{(3)}$

α	ν^*	ν^{**}	Asymp.	$z_{40}^{(3)}(\alpha)$
0.1	3.15	3.29	3.60	3.41
0.05	3.41	3.54	4.05	3.71
0.01	3.93	4.04	5.06	4.26

Table III. Change in the mean (variance is unknown and constant)

	$z_{40}^{(1)}$	u^{**}	Asymp.	\hat{k}	Before	After
February	0.1	0.1		15	3.43	4.06
March	0.05	0.01	0.1	23	3.26	4.02
April	0.1	0.1		15	3.21	3.85
May	0.05	0.05		26	3.92	4.37
November	0.05	0.05		26	2.94	3.56

Table IV. Change in the variance (mean is unknown and constant)

	$z_{40}^{(2)}$	u^{**}	Asymp.	\hat{k}	Before	After
February	0.05	0.01	0.05	3	0.0023	0.6230
March		0.1		27	0.7563	0.1677
April	0.01	0.01	0.05	37	0.5391	0.0002
November	0.01	0.01	0.05	32	0.5402	0.0227

Table V. Change in the mean and/or the variance

	$z_{40}^{(2)}$	ν^{**}	Asymp.	\hat{k}	Mean before	Mean after	Variance before	Variance after
February	0.05	0.05	0.1	3	3.47	3.85	0.0020	0.4859
March	0.05	0.05	0.1	23	3.26	4.02	0.4927	0.2492
April	0.01	0.01	0.05	37	3.59	3.81	0.4945	0.0003
November	0.05	0.05	0.05	32	3.11	3.38	0.4874	0.0212

We computed the values of $T_{40}^{(1)}$, $T_{40}^{(2)}$ and $T_{40}^{(3)}$ for each month. The results are summarized in Tables III to V. Only those months are listed where significant changes were found. The smallest values of $\alpha = 0.1$, 0.05 and 0.01 which are still significant are also given. \hat{k} is the time where $-2 \log$ of the likelihood function reaches its maximum. According to Gombay and Horváth (1996b), \hat{k} is a consistent estimator for the time of change. It is clear from Tables III and V that the mean increased after the change. However, the variance decreased after the change except in February. For comparison the graphs of $-2 \log$ of the likelihood ratios for March and April are given in Figures 1–6. Removing the third observation from the April data, the change in the variance will disappear. It looks like an early outlier was picked up as a possible change in the variance. The study gave strong evidence for the change in both parameters in April and November. It is also very likely that the mean increased in February and March while the variance remained stable in these months. Also, probably the mean increased first and after that the variance decreased.

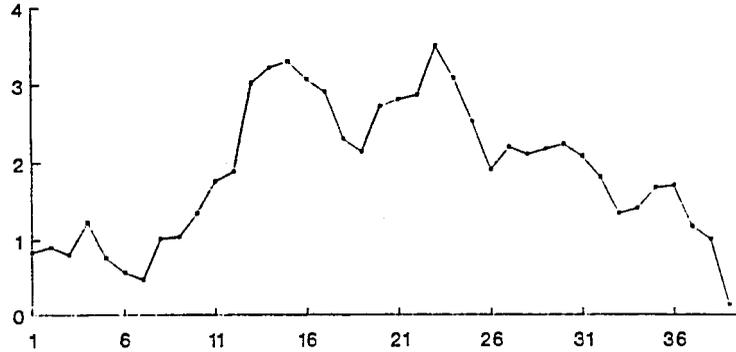


Figure 1. The graph of $-2 \log \Lambda_k^{(1)}$ for March



Figure 2. The graph of $-2 \log \Lambda_k^{(2)}$ for March

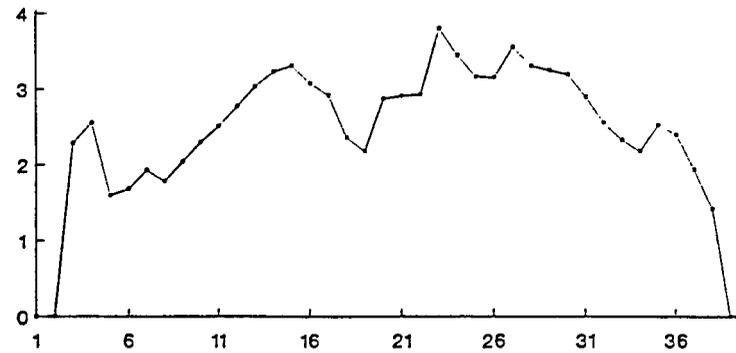
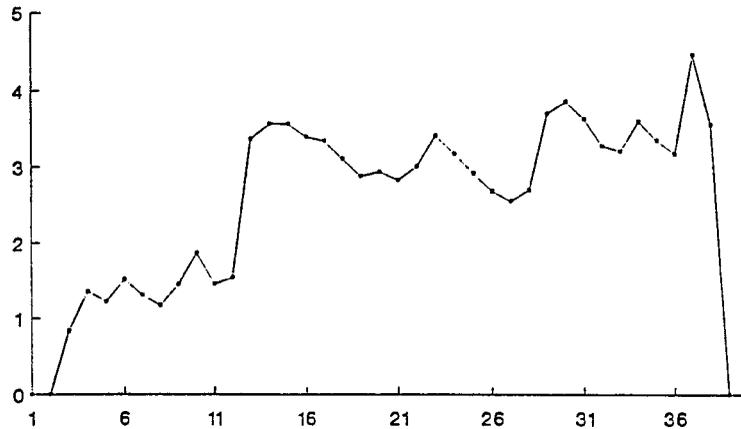
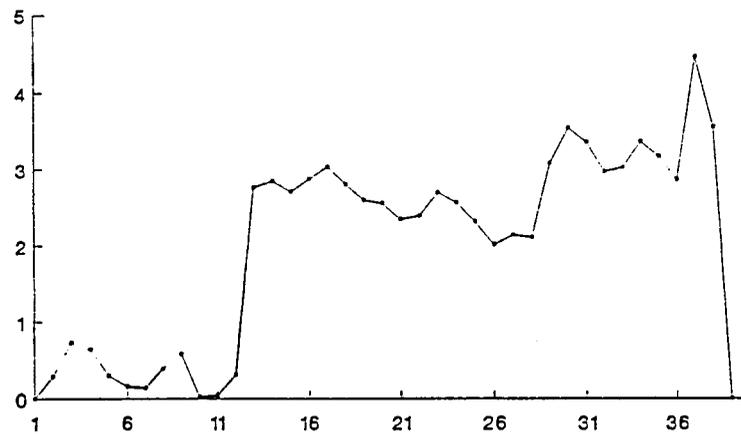
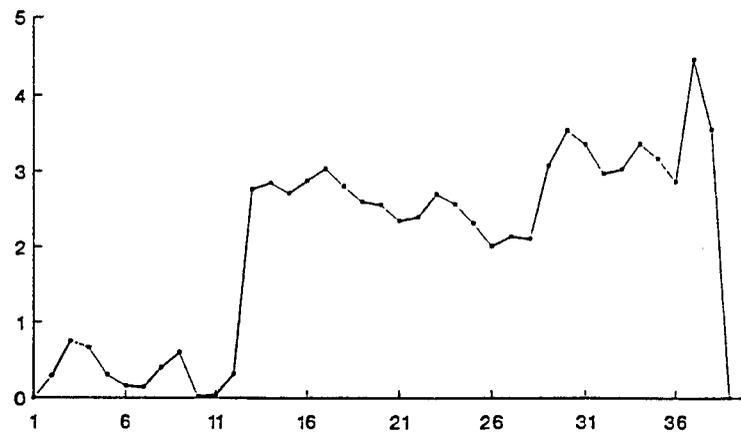


Figure 3. The graph of $-2 \log \Lambda_k^{(3)}$ for March

3. DEPENDENT OBSERVATIONS

Instead of investigating the water discharges for each month separately, we consider the full sequence. It is very unlikely that the consecutive months are independent, so we need a new model in which the possible dependence is incorporated. The data are also periodic. Let

Figure 4. The graph of $-2 \log \Lambda_k^{(1)}$ for AprilFigure 5. The graph of $-2 \log \Lambda_k^{(2)}$ for AprilFigure 6. The graph of $-2 \log \Lambda_k^{(3)}$ for April

and

$$\tilde{B}(k) = \begin{pmatrix} \tilde{\sigma}_1^2 & \tilde{v}_1 & 0 & \dots & \tilde{v}_d \\ \tilde{v}_1 & \tilde{\sigma}_2^2 & \tilde{v}_2 & \dots & \\ \vdots & & & & \\ & & & & \tilde{\sigma}_d^2 \tilde{v}_{d-1} \\ \tilde{v}_d & & & & \tilde{\sigma}_d^2 \end{pmatrix}$$

with

$$\hat{\sigma}_j^2 = \hat{\sigma}_j^2(k) = \frac{1}{k/d} \sum_{i \in I(j,k)} (X_i - \hat{\mu}_j(k))^2$$

$$\tilde{\sigma}_j^2 = \tilde{\sigma}_j^2(k) = \frac{1}{(n-k)/d} \sum_{i \in I(j,k)} (X_i - \tilde{\mu}_j(k))^2$$

$1 \leq j \leq d$ and

$$\hat{v}_1 = \hat{v}_1(k) = \frac{1}{k/d} \sum_{\substack{i \in I(1,k) \\ i+1 \leq k}} (X_i - \hat{\mu}_1(k))(X_{i+1} - \hat{\mu}_2(k))$$

$$\tilde{v}_1 = \tilde{v}_1(k) = \frac{1}{(n-k)/d} \sum_{i \in L(1,k)} (X_i - \tilde{\mu}_1(k))(X_{i+1} - \tilde{\mu}_2(k))$$

and $\tilde{v}_2, \dots, \hat{v}_d, \tilde{v}_d$ are defined in similar ways. We reject the 'no change in the mean' null hypothesis if

$$T_n = \max_{1 \leq k < n} (\hat{\mu}_1(k) - \tilde{\mu}_1(k), \dots, \hat{\mu}_d(k) - \tilde{\mu}_d(k))$$

$$\times \left(\frac{1}{k/d} \tilde{B}(k) + \frac{1}{(n-k)/d} \hat{B}(k) \right)^{-1} \begin{pmatrix} \hat{\mu}_1(k) - \tilde{\mu}_1(k) \\ \vdots \\ \hat{\mu}_d(k) - \tilde{\mu}_d(k) \end{pmatrix}$$

is large. Following the method of Horváth and Shao (1995) one can prove that

$$\lim_{n \rightarrow \infty} P \left\{ \left(2 \log \log \frac{n}{d} \right)^{1/2} T_n^{1/2} \leq x + 2 \log \log \frac{n}{d} + \frac{d}{2} \log \log \log \frac{n}{d} - \log \Gamma(d/2) \right\} = \exp(-2e^{-x}) \quad (7)$$

where $\Gamma(t)$ stands for the gamma function.

In our example $n = 480$ and $d = 12$. Using (7) we get that a significant change occurred at $\alpha = 0.1$ significant level.

ACKNOWLEDGEMENTS

We wish to thank Professor Daniela Jarušková for kindly providing the data on Načetínský. The research of Edit Gombay was partially supported by an NSERC Canada operating grant.

Research of Lajos Horváth was partially supported by the NATO Scientific and Environmental Division (CRG 960503).

REFERENCES

- Csörgő, M. and Horváth, L. (1997). *Limit Theorems in Change-Point Analysis*, Wiley, Chichester.
- Gombay, E. and Horváth, L. (1996a). 'On the rate of approximations for maximum likelihood test in change point models'. *Journal of Multivariate Analysis*, **56**, 120–152.
- Gombay, E. and Horváth, L. (1996b). 'Approximations for the time of change and the power function in change-point models'. *Journal of Statistical Planning and Inference*, **52**, 43–66.
- Horváth, L. and Shao, Q.-M. (1995). 'Limit theorems for the union-intersection test'. *Journal of Statistical Planning and Inference*, **44**, 133–148.
- Jarušková, D. (1994). 'Applications of change-point detection to ecology'. Preprint.
- Vostrikova, L. Ju. (1981). 'Detection of 'disorder' in a Wiener process'. *Theory of Probability and its Applications*, **26**, 356–362.
- Yao, Y.-C. and Davis, R. A. (1984). 'The asymptotic behavior of the likelihood ratio statistic for testing a shift in mean in a sequence of independent normal variates'. *Sankhya Series A*, **48**, 339–353.