MATEMATIKA, 2013, Volume 29, Number 2, 149–158
©Department of Mathematics, UTM.

Checking the Dependency of *E-coli* Data with Extremal Index Using Block and Run Estimators

[1]**Mohd Bakri Adam and** [2]**Noor Fadhilah Mohd Ramlan**

[1]Department of Mathematics,
Universiti Putra Malaysia

[2]Universiti Teknologi MARA,
Cawangan Machang, Malaysia

e-mail: [1] bakri@upm.edu.my

Abstract The extremal index characterizes the degree of local dependence in the extremes of a stationary time series. In this paper, we discuss an alternative interpretation of extremal index as a ratio of the limiting expected value of two random variable defined by extreme levels and a partition of the stationary sequence into blocks. We use the run and the block definition for the cluster to analyse the clustering to find extremal index. These experiments highlight the importance of block size selection. The practical implications are examined through the *E-coli* data from rivers in Selangor.

Keywords Block and Run Estimators; Dependency; *E-coli*; Extremal Index

2010 Mathematics Subject Classification 60G70, 62G32, 58E15.

1 Introduction

The extremal index, θ, is an important parameter for extending extreme value theory results from independent and identically distributed sequences to stationary sequences. For discrete-time continuous-valued stationary time series sequence, we consider the impact of dependence on the extreme values. With practical applications, it is usual to assume a condition that limits the extent of long range dependence at extreme levels, so that the events $X_i > u$ and $X_j > u$ are approximately independent, provided u is high enough and time points i and j have a large separation. Only processes with any form of short range dependence for which at long lags the extreme are independent are consider. In other words, extreme events are close to independent at times that are far enough apart. Many stationary series satisfy this property.

In this paper, we are going to implement the extreme index to check the dependency of *E-coli* bacteria in Selangor rivers. We found that within our knowledge that no researcher has been carried out the relationship of existence of *E-coli* in Malaysia rivers. We are aim to look at the influence of the group of *E-coli* by the existing of other group of *E-coli* by estimating the extremal indexes using two methods i.e. block selection and run methods.

A condition that makes precise the notion of extreme events being near-independent if they sufficiently distant in time. Let $X_1, \ldots, X_n$ be a stationary sequence of random variables satisfying the long-range asymptotic independence condition $D(u_n)$ of Leadbetter [1] then

$$Pr\{\max(X_1, \ldots, X_n) \leq u_n\} \to \{G(x)\}^\theta, \qquad \text{as } n \to \infty \tag{1}$$

where u_n is a threshold and $u_n = a_n x + b_n$ with $a_n > 0$ and b_n selected to ensure that the limit distribution is non-degenerate, and $G(x)$ is the limit distribution of

$$\Pr\{\max(X_1^*, \ldots, X_n^*) \leq u_n\}, \tag{2}$$

where the sequence of random variables $X_1^*, \ldots, X_n^*$ are independent but with the same univariate marginal distribution as $X_1, \ldots, X_n$ [2].

The extremal index gives a measure of the short range dependence exhibited by the extremes of a process. In particular case, the tendency of the extremes to occur in cluster can be indicated. A particularly important special case is when $\theta = 1$ which, as well as occurring for independent sequences, arises for a broad range of dependent sequences for which the limiting clusters occur as single values. For example, all weakly mixing Gaussian processes have $\theta = 1$, see [1] and [3]. When $D(u_n)$ condition is satisfied then $G(x)$ condition also being satisfied leading to value of $\theta = 1$. This result shows that θ is the parameter measure of short range dependence in extreme values as the asymptotic behavior of $\max(X_1, \ldots, X_n)$ and $\max(X_1^*, \ldots, X_n^*)$ is identical even though $X_1, \ldots, X_n$ may be dependent.

There are at least two ways of analyzing the clustering. We might define clustering with respect to a block of some period or with respect to run of exceedances above the threshold. Each definition of a cluster reveals a different aspect of the study. The second one gives us information on the stable behaviour of estimated extremal index above the threshold. In case of a large clustering it indicates none rapidly varying estimated extremal index series around the threshold. A significant change of the cluster index would show a change of this stability of longer or shorter periods of study.

The limiting mean cluster size where a cluster is defined as the set of exceedances of the threshold, u_n is clarified by [2] and [4] . Given that at least one exceedance occurs in the block where the sample of size n is divided into blocks of length r_n with $r_n = o(n)$. With this definition of a cluster, the cluster size distribution π_n is defined as

$$\pi_n(j; u_n, r_n) = \Pr\left\{\sum_{i=1}^{r_n} I(X_i > u_n) = j | \sum_{i=1}^{r_n} I(X_i > u_n) > 0\right\}, \text{ for } j = 1, \ldots, r_n \quad (3)$$

with

$$I(X_i > u_n) = \begin{cases} 1 & \text{if } X_i > u_n, \\ 0 & \text{otherwise} \end{cases}$$

where I is the indicator function. Later, Hsing and Leadbetter [12] showed that

$$\theta^{-1} = \lim_{n\to\infty} \sum_{j=1}^{r_n} j\pi_n(j; u_n, r_n) \quad (4)$$

is the limiting mean cluster size [3]. An alternative representation for θ was given by O'Brien [5] who showed that

$$\theta = \lim_{n\to\infty} \Pr(X_i \leq u_n, 2 \leq i | X_1 > u_n) \quad (5)$$

where $r_n = o(n)$ [3].

These suggest that estimating θ and identifying independent clusters are fundamentally important for statistical applications for stationary processes. The mean cluster size is estimated by the sample average cluster size. It has been developed based on two characterizations for θ. Both estimators can be expressed as

$$\hat{\theta}_n = \frac{C_n(u_n)}{N_n(u_n)}, \text{ as } n \to \infty \quad (6)$$

where $N_n(u_n)$ is the total number of exceedances of a high threshold, u_n by $X_1, \ldots, X_n$ and $C_n(u_n)$ is the number of independent clusters above threshold, u_n. Thus, estimating θ is equivalent to identifying independent clusters.

2 Estimation of the Extremal Index

Clustering of such extreme events can be measured by the extremal index. It can be interpreted as the inverse of the mean number of extreme events in a cluster. The extremal index has a value of unity in independently distributed data. If the data are serially dependent but show no tendency to give clusters of extreme values then this might suggest that the underlying process has extremal index, $\theta = 1$. If extreme values display some clustering, it suggest that an extremal index, $\theta < 1$, which means the assumption of independent excess losses is less satisfactory.

There are two methods to estimate the extremal index, θ. These two methods estimate a mean number of observations in a cluster. Using either of these two methods requires a definition of a cluster, so that distinct clusters maybe identified and separated, and of an associated de-clustering parameter, r. The first method divides the data into approximately k_n blocks of length r_n, where $n \approx k_n r_n$. Each block is treated as one cluster. On each block, compute the maximum,

$$M_r^{(i)} = \max_{i=1,\ldots,k} (X_{(i-1)r+i}, \ldots, X_{ir}). \tag{7}$$

The extremal index can be estimated by the sample analogue of

$$r^{-1} \left[\frac{\Pr(M_r^{(i)} > u_n)}{\Pr(X_i > u_n)} \right] \tag{8}$$

or

$$\hat{\theta} = \frac{\sum_{i=1}^{k} I(M_r^{(i)} > u_n)}{\sum_{i=1}^{k} I(X_i > u_n)} = \frac{n_B}{n} \tag{9}$$

where I is the indicator function, u_n is the threshold, n_B is the number of blocks showing at least one value in excess of the threshold, and n is the total number of exceedances of the threshold on the sample. The interpretation of extremal index, θ as the inverse of the mean cluster size is then natural, where each block showing at least one exceedance is defined to be a cluster.

The second method based on a de-clustering method is that of run, in which a cluster is defined as the fist cluster starts with the firrst exceedance of threshold, u_n and stops as soon as there is a value below threshold, u_n. The second cluster starts with the next exceedance of threshold, u_n and so on. Define a set of clusters by the condition $C_i : X_i > u_n, i+1 \leq u_n, \cdots, X_{i+r} \leq u_n$. Then the number of clusters is the number of times that the condition holds on the sample,

$$n_C = \sum_{I(C_i)}^{T-r}. \tag{10}$$

The estimator of the extremal index is

$$\hat{\theta} = \frac{\sum_{i=1}^{T-r} I(C_i)}{\sum_{i=1}^{k} I(X_i > u)} = \frac{n_C}{n} \tag{11}$$

We call this the run estimator. For each of these estimators we must have $r = r(n)$ such that $r \to \infty$.

In applications, the sample size is always finite. If it has overlapping between two successive subintervals, a run of exceedances is split up into two clusters. Run of exceedances within a subinterval are joined to give a block of exceedances. Therefore, it is obvious that the definition of cluster will be influence by the estimation of θ.

Figure 1 shows portion the daily maxima of E-coli level recorded by all rivers in Selangor From the figure we count the point which is above the threshold, $n = 9$. Then, we count the number of clusters, $n_C = 7$. Hence, $\hat{\theta} = \frac{7}{9} = 0.77778$. The $\hat{\theta}$ being close to 1 indicates to closely independence of clustering. So, we can say that it is gives closely independent results.

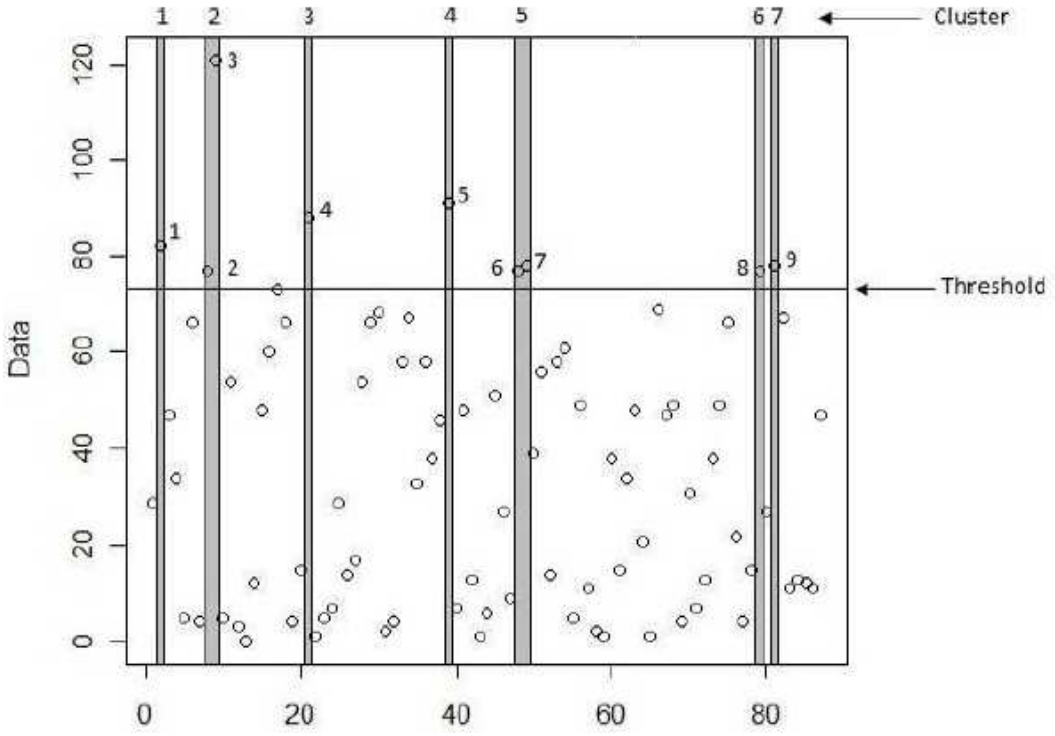

Figure 1: Portion the Daily Maxima of E-coli Level Recorded by All Rivers in Selangor

The extremal index measures the strength of the dependence. The value of $\theta = 0$ corresponds to a long memory sequence, $0 < \theta < 1$ a short memory sequence, and $\theta = 1$ is a no memory. If $\theta > 0$, then the dependence is weak so that M_n from equation (1) can be normalized with the normalization which is appropriate for the maximum of n independent and identically distributed random variables from F, whereas if $\theta = 0$, then the dependence is so strong that a different normalization is called for. Thus the nature of long memory sequences is very different from that of short or no memory sequences.

Consistency of the block and run estimators was established by [1] and [6] respectively,

and consistency and asymptotic normality are established under different sets of conditions by [7] and [8] for the block and run estimator [9].

The theoretical properties of the blocks estimator have been studied by [10], [11]. It is a natural starting point for any statistical study because probabilistic analyses of the extremal index [1] and [12] used the same definition. However, the runs estimator seems more natural from a statistical point of view. For example, Smith [13] used precisely this method of separating clusters, calling r_n the cluster interval. Nandagopalan [14] and Leadbetter [15] have used a version of the runs estimator with $r_n = 1$ and Nandagopalan [14] derived theoretical properties for it, but in general there seems no reason to confine ourselves to $r_n = 1$ and a more general estimator is desirable [16].

2.1 Selection of Threshold

The estimation of extremal index, θ is also influenced by the choice of threshold, u_n. Selecting a rather large threshold, u_n would mean that, there are only very few exceedances (with large probability). We know from the theory that the boundary threshold, u_n is chosen such that $n(1 - F(u_n)) \to \tau > 0$. Since the number of exceedances has asymptotically a Poisson distribution, there is a positive probability that no exceedances above the boundary threshold, u_n occurs in which case extremal index, θ cannot be estimated.

For some boundaries threshold, u_n we have $n = 1$ and therefore $n_C = 1$ which gives $\hat{\theta}_n = 1$. For example, if threshold, u_n is such that $n = 2$ then $\hat{\theta}_n = 1$ or $= \frac{1}{2}$ depending on n_C, similarly for $n > 2$. This implies that the estimator $\hat{\theta}_n$ cannot be consistent for every $\theta \in (0, 1]$, if threshold, u_n is chosen in such a way that n is fixed.

Therefore it was proposed to use $v_n < u_n$ such that n tends to ∞, i.e. $n(1-F(v_n)) \to \infty$ (slowly). Hsing [11] and Nandagopalan [14] proved that, with this choice the estimator $\hat{\theta}_n$ of θ is consistent, whether one use the run or the block definition for clusters. Leadbetter and Rootzen [15] proposed to use v_n approximately 5 percent, 7.5 percent or 10 percent. Hsing [11] showed using the block definition for clusters and the above choice of v_n that under certain additional conditions the estimator $\hat{\theta}_n$ has asymptotically a normal distribution.

2.2 Choice of De-clustering Parameter, *r*

In a particular choice of threshold, u_n the separation of extreme events into clusters is likely to be sensitive. To overcome these deficiencies, consider a cluster to be active until r consecutive values fall below the threshold, u_n for some pre-specified value of r. For example Figure 2 shows the effect of different choices of r on cluster identification. With $r = 1$ seven clusters are obtained. With $r = 2$ just four clusters are obtained. The choice of r requires care. If a value of r is too small, it will lead to the problem of independence being unrealistic for nearby clusters. If a value of r is too large, it will lead to a concatenation of clusters that could reasonably have been considered as independent and therefore to a loss of valuable data.

Under rather general conditions, there exists some finite r such that

$$\lim_{x \to \infty} \Pr(V_{1 \le j \le r} \xi_j \le x | \xi_0 > x) = \theta \tag{12}$$

If equation (12) holds for some $r = m$ then it holds for all $r \ge m$. Therefore, it is clear that if $\{\xi_j\}$ is $m-$dependent, then equation (12) holds with $r = m$. In many cases where the

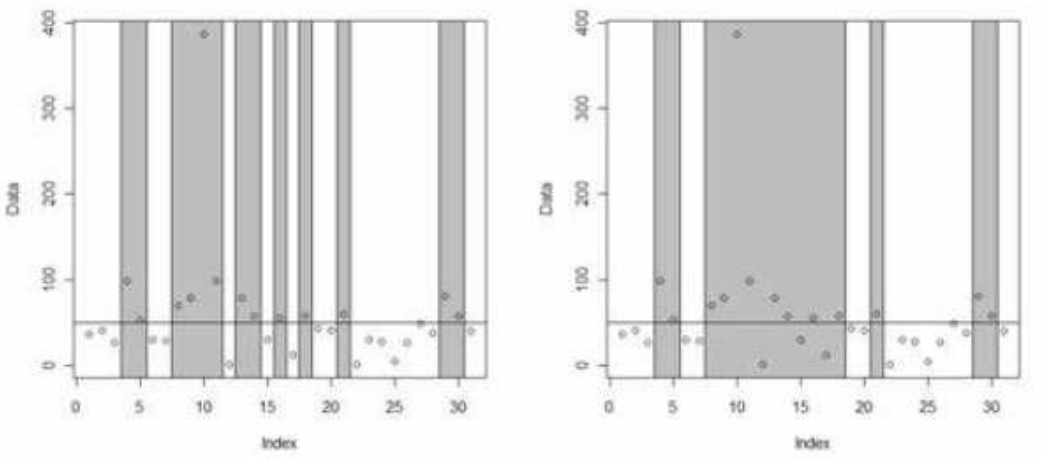

Figure 2: Portion the Daily Maxima of E-coli Recorded for Selected Rivers in Selangor with Two Possible Cluster Groupings

range of dependence is infinite, equation (12) actually hold with $r = 1$. See [7], [17], [5], [16] and [18].

If we are given data $\xi_1, \ldots, \xi_n$ we can choose some r which reasonably covers the range of dependence. The estimated θ is

$$\hat{\theta} = \left(\frac{\sum_{1 \le j \le n-r} I(\xi_j > x)}{\sum_{1 \le j \le n-r} I(\xi_j > x \le V_{j+1 \le i \le j+r} \xi_i)} \right) \tag{13}$$

where x is a properly chosen large value. This is known as the runs estimate of the extremal index. A good way of interpreting this procedure is that, first identify clusters of exceedances in the data as those that are separated by at least r non-exceedances and then estimate θ by the reciprocal of the average number of exceedances in such clusters.

3 Applications

In this section we will apply both methods which are the run and the block definition for the cluster to the real data. From the data, we observe that, the numbers of *E-coli* (in part per thousand) were recorded between 2 to 13 times per day, approximately 8 days per month from 35 rivers in Selangor from January 2007 through out December 2007. Also recorded 40 others parameter such that temperature, pH of the river water and weather.

The quality of the data can be described as generally quite good. However, for data of this size, it is inevitable to have missing observations and round-off and other types of error. For our purpose we use the data segment of which the quality exceptional, namely the daily maxima of *E-coli* level recorded by rivers in Selangor from January 2007 to December 2007. We obtained this data from Malaysia's Department of Environment.

Our major concern here is to illustrate how to find the estimator of the extremal index of the *E-coli* to predict wether the data is dependent or independent and also bring up some

issues that need to be addressed in statistical extreme value theory for dependent data.

The number of *E-coli* was measured 545 times at rivers in Selangor for 12 months. From the data, we only use the maximum level of *E-coli* per day. Then, we will get $n = 80$ numbers of measurement for *E-coli* to represent this study.

In order to obtain approximately 10 percent, 7.5 percent or 5 percent for $\frac{N}{n}$, v_n should be 397, 398 and 399 (in thousand) *E-coli*, respectively. Thus $N = 8$; 6 and 4, respectively. Only the values larger than 397 (in thousand) *E-coli* were recorded. See Table 1.

To analyse the behaviour of the estimator with respect to v_n, we continuously change the value of v_n, from 397 up to the maximal possible value 590.

Table 1: The Values Larger Than 397 (in Thousand) *E-coli*

Number of Measurement, i	*E-coli*
36	540
143	590
299	490
309	399
315	510
442	398
486	399
530	398

3.1 Clusters by Run

In our analysis we first use the run definition of clusters and later the block definition. From the Figure 3, obviously, θ_n is a piecewise constant of v_n. But we cannot say that the estimator is monotonically increasing. We can observe that θ_n has tendency to increase in v_n.

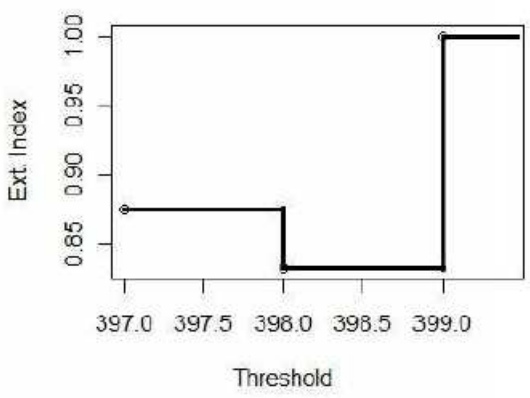

Figure 3: Estimation of θ by $\hat{\theta}_n$ using the Run Definition for Clusters

From Table 2, we observed that the estimation of θ for $v_n = 399$ is 1. We can say that it is gives independent results for this data. This fact can also be deduced from Figure 3. There is large enough interval of v_n where $\hat{\theta}_n$ is stable at 1, which could be used as creation.

Table 2: Estimation of θ Based on v_n using the Run Definition for Clusters

v_n	N	Z	$\hat{\theta}_n$
397	8	7	0.875
398	6	5	0.833
399	4	4	1.000
510	2	2	1.000
540	1	1	1.000

3.2 Clusters by blocks

When we use the block definition for the clusters, we carry out the analysis of the same data using subintervals of length 5 and 10 as a blocks, see Table 3, again for all levels v_n larger than 397 to indicate the dependency of θ_n.

We deduce from these case, that the estimates of θ are equal for the three cluster definitions if $v_n \leq 399$. The same is true for the cluster size distributions gives a good reason for selecting $v_n = 399$ as an appropriate level for the estimation. The $\hat{\theta}_n = 0.75$ is selected as the estimated for θ. θ_n being close to 1 indicates independence of clustering. So, we also can say that it is gives independent results for this data. See Figure 4.

Table 3: Estimation of θ Based on v_n using the Block Definition for Clusters

v_n	N	Z	$\hat{\theta}_n$	v_n	N	Z	$\hat{\theta}_n$
397	8	7	0.750	397	8	7	0.500
398	6	5	0.667	398	6	5	0.667
399	4	3	0.750	399	4	3	0.750
510	2	2	1.000	510	2	2	1.000
540	1	1	1.000	540	1	1	1.000

4 Discussion and Conclusion

This research focuses on the estimation of the extremal index for *E-coli* data in rivers at Selangor. We found that each *E-coli* observation is closely independent to others *E-coli* in all Selangor rivers that have been selected. This may be due to different climatic factors or sources of the different findings from the surrounding water.

We also can make the highlight that the importance of block size selection. From this research, when we are using the run definition for the cluster to analyse the clustering to

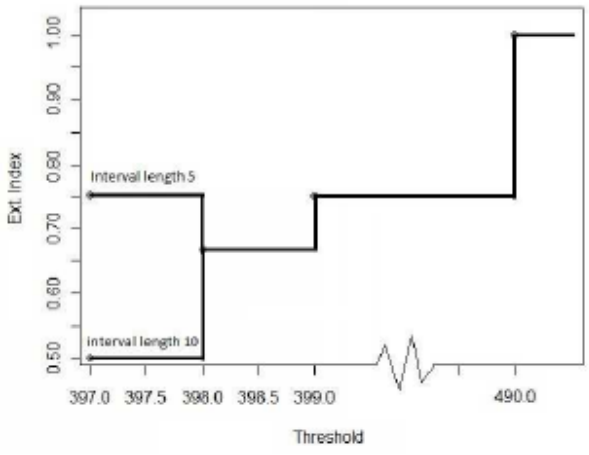

Figure 4: Estimation of θ by $\hat{\theta}_n$ using the Block Definition for Clusters for Interval of Length 5 and 10

find extremal index, we only can see at the interval of v_n where θ_n is stable, which could be used as creation to select the estimated for θ. By using the block definition for the cluster, we can find the appropriate value of v_n by looking at the three cluster definitions and then can find the estimated for θ. There are two number of issues that need to be addressed i.e. more studies are required on the extreme of non-stationary models and models with seasonal features and in the estimation of the extremal index, further investigation is required to enhance the understanding of how to choose the tuning parameters.

References

[1] Leadbetter, M. Extremes and local dependence in stationary sequences. *Zeitschrift fur Wahrscheinlichkeitstheorie und Verwandte Gebiete*. 1983. 65: 291-306.

[2] Laurini, F. and Tawn, J. *New Estimators for the Extremal Index and Other Cluster Characteristic*. New York: Kluwer Academic Publishers. 2003. (6): 189-211.

[3] Ancona-Navarrete, M. and Tawn. J. *A Comparison of Methods for Estimating the Extremal Index*. New York: Kluwer Academic Publishers (3): 5-38. 2000.

[4] Leadbetter and Rootzen, H. Extremal theory for stochastic proceses. *Ann. Probab.* 1988. 16: 431-478.

[5] O'Brien, G. Extreme values for stationary and Markov chains with applications. *Ann. Probab.* 1987. 15: 281-291.

[6] O'Brien, G. The maximum term of uniformly mixing stationary processes. *Z. Wahrsch. v. Geb.* 1974. 30: 57-63.

[7] Hsing, T. On tail index estimation using dependent data. *Annals of Statistics*. 1991. 19: 1547-1569.

[8] Weissman, I. and Novak, S. On blocks and runs estimators of the extremal index. *Journal of Statistical Planning and Inference.* 1998. 66: 281-288.

[9] Galbraith, J. and Zernov, S. *Extremes and Extreme Dependence in the NASDAQ and S and P 500 Composite Indexes.* ES World Congress. 2005.

[10] Hsing, T. *Estimating the Extremal Index Under M-dependence and Tail Balancing.* Department of Statistics, Texas A and M University: Ph.D. Thesis. 1990.

[11] Hsing, T. Estimating the parameters of rare events. *Stoch.* 1991. 31: 117-139.

[12] Hsing, T. and Leadbetter, M. On the exceedance point process for a stationary sequence. *Prob. Theory and Related Fields.* 1988. 78: 97-112.

[13] Smith, R. Extreme value analysis of environmental time series: an application to trend detection in ground-level ozone. *Statistical Science.* 1989. 4: 367-393.

[14] Nandagopalan, S. *Multivariate Extremes and Estimation of the Extremal Index.* University of North Carolina, Chapel Hill: Ph.D. Thesis. 1990.

[15] Leadbetter, M. R. and Rootzen, H. On clustering of high values in statistically stationary series. In *Proc. 4th Int. Meet. Statistical Climatology.* 1989. 217-222.

[16] Smith, R. and Weissman, I. *Estimating the Extremal Index.* New York: Springer. 1991.

[17] Hsing, T. Extremal index estimation for a weakly dependent stationary sequence. *Annals of Statistics.* 1993. 21: 2043-2071.

[18] Reiss, R. D. and Thomas, M. *Statistical Analysis of Extreme Index.* Switzerland: Birkhauser. 1991.