International Journal of Reliability and Applications
Vol. 7, No. 2, pp.101-110, 2006

Robust Most Significant Periods of Developments In Time Dominated Data

F. Aboukalam[1]
Department of Statistics and O. R., College of Sciences
P. O. Box 2455, King Saud University, Riyadh 11451, Saudi Arabia

Abstract. Let E be a set of n quantitative observations under the time control. The interval of time is to be split into several subintervals such that the observations in each subinterval are almost similar, whereas the observations between the subintervals are very dissimilar. The corresponding time-subintervals become periods or phases of the development that exist in the underlying phenomenon. Aboukalam(2005) proposes a robust solution based on some initial subintervals and a technique for combining any two successive groups in that starter using a t-test under a fixed significant level (α). The inconvenience is that; the technique reliability is not released from the level α which must not be defined apart from the number of the periods that is, in its turn, unknown. To avoid this, we propose what so called; most significant periods solution. The new technique constructs its own initial subintervals and uses another way for combining the groups. However, the way of determining and treating outliers has not changed. This paper conducts many empirical simulations using different possible time dominated data in order to illustrate the reliability of the proposed technique. Finally, we apply both techniques on some real time dominated data to explain the advantage of the proposal.

Key Words: Clustering, outlier, time dominated data, periods.

1. INTRODUCTION

Let $\boldsymbol{E}$ be a set of n one-dimensional quantitative observations under the domination of time. The idea, as in cluster analysis, is to divide $\boldsymbol{E}$ into some groups, such that there is homogeneity within the groups and heterogeneity between the groups. For the specific time dominated data, it is quite sensible to consider the time intervals of the corresponding groups as periods or phases of

[1] My first name was changed from (M. A. Fayez) in to (Fayez)
Corresponding author. E-mail: abokalam@ksu.edu.sa

the development in the data. Often, the true number of the periods, k_0, is not known, and we should look for a suitable estimate $\hat{k}_0$. In the literature, Hartigan(1975) uses Fisher's technique to divide the data into several possible numbers of optimal groups, and proposes a specific F-test to optimize one of the different divisions. Aboukalam(2005) shows how Hartigan's test reliability is very low, and proposes an alternative technique with a way for determining and treating outliers. The technique divides $\boldsymbol{E}$ sequentially into ascending numbers of optimal groups up to a number, $\boldsymbol{M}$, chosen by an automatic stopping rule. As $\boldsymbol{M}$ is empirically always greater than k_0, Aboukalam proposes to recombine any two successive groups i and $i+1$ if the t-test cannot significantly reject the hypothesis $H_0 : \mu_i = \mu_{i+1}$ under some level α. To make the results considerable, any group should not be dwarfish; namely, whose size is lesser than 5, say. If it is so, the rule goes back to the nearest free dwarfish division with $\boldsymbol{L}$-groups, say. The simulation shows the inequality $L \geq k_0$ is true 99.5%, which, again, justifies the fusion approach. Eventually, Aboukalam could not fairly avoid the nuisance between the technique reliability and the level α except α is taken very small.

We propose here the robust most significant solution as an alternative to the solution in Aboukalam(2005). The solution will not give any role to the level α and, as a result, there will be some increases in the reliability. The stopping rule and the way for determining and treating outliers will not be changed. However, the initial free dwarfish $\boldsymbol{L}$-division will not be the same. Precisely, each dwarfish group is combined with the neighbor that make the within sum of squared errors lesser. Eventually, simulation experiments and an example on real data will highlight the differences with the solution in Aboukalam(2005).

2. A STOPPING RULE

Fisher's algorithm divides the set $\boldsymbol{E}$ into a given number k of exact optimal groups. The algorithm assigns the observations to k-groups in which the within group sum of squares, $W(n, k)$, is exactly minimized. Fisher's algorithm successively gets optimal l -groups for I objects, where $2 \leq l \leq k-1$ and $l \leq I \leq n$, and it builds on them to find optimal $(l+1)$-groups, and so on until the aimed optimal k-groups is achieved. This is called a dynamic programming procedure; see Bellman and Dreyfus(1962).

Many simulations on time dominated data that have all potential k_0-phases showed that the principal decreases in $W(n, k)$ come when $k \leq k_0$ and later decreases are relatively minor. Hence, an objective choice of the upper bound, $\boldsymbol{M}$, for k may be as follows: wait for the first three successive times where in each of which the value of $|W(n,k)/W(n,k+1)-1|$ does not exceed a small figure, like 0.10, and denote $\boldsymbol{M}$ to the first k values.

The simulation in Sec 3 shows that k_0 cannot be greater than M and the average of the difference (M-k_0) is equal to 1.4, which indicates the high rapidity of the rule.

3. SIMULATION EXPERIMENTS

Let $\boldsymbol{E}$ be artificially built up to consist of n =100 normal observations, and is divided into k_0 -periods each of which has size n_i, mean μ_i and unit s.d $\sigma = 1$. The way of selection of k_0, μ_i, n_i will be random in order to make the phases free of any possible bias in their shapes and masses. The experiment conducts the cases k_0 =3, 4, 5, 6 and 7, and for each k_0 the means $\{\mu_i, i = 1,..,k_0\}$ are randomly selected from the set {0, 5, 10, 15, 20, 25, 30} with repeat providing that any two consecutive means must not be equal. The size of period i is given the value; $n_i = n \times w_i / \sum_{j=1}^{k_0} w_j$, where w_j is randomly selected with repeat from the set of weights {8,12, 16, 20, 24, 28, 32}. When the data $\boldsymbol{E}$ should have some outliers (***OE***), we repeat the simulation after we change the unit s.d $\sigma = 1$ of 10 observations selected at random by $\sigma = 15$.

For each data, the simulation computes the numbers $\boldsymbol{M}$ and L in order to notice the rapidity of the stopping rule and to justify the fusion approach. Moreover, in order to asses the technique reliability; we should find the estimate $\hat{k}_0$ and, if the estimation it is correct (*i.e.*$\hat{k}_0 = k_0$) we should look for the number of objects that missed their actual groups (*NOMG*).

We repeat the simulation 500 times, and in the end we compute the probability $\Pr(\hat{k}_0 = k_0)$ and the average of *NOMG*'s values, say *ANOMG*, among the correct runs.

4. ALTERNATIVE TECHNIQUE

4.1- Most Significant Periods – Solution

The technique builds up its initial free dwarfish $\boldsymbol{L}$-division from the M-division by combining each dwarfish group with the neighbor that make the within sum of squared errors lesser. The simulation shows the inequality $L \geq k_0$ is 99.6% true, and in those cases the average of the differences $(L - k_0) = 0.58$. This justifies the fusion approach in the technique to retrieve the real k_0-periods.

The technique abbreviates the term x_{ij} to the element $j = 1,..., n_i$ where n_i is the size of group $i = 1,..., L$. Moreover, we assume that x_{ij} follows the normal law

$N(\mu_i, \sigma^2)$ where σ^2 is a common variance, and that the initial L-division rejects the hypothesis: $H_0 : \mu_1 = ... = \mu_L$ under a risk π_L. The risk incurred by the L-division is the probability: $\pi_L = 1 - G(F_L, L-1, n-L)$ where F_L is the F-ratio:

$$F_L = \frac{\sum_{i=1}^{L} n_i (x_{i.} - x_{..})^2 / (L-1)}{\sum_{i=1}^{L} \sum_{j=1}^{n_i} (x_{ij} - x_{i.})^2 / (n-L)}$$

and G is the c. d. f of F-distribution with d.f $(L-1)$ and $(n-L)$. In the next step, we combine each two successive groups in the underlying L-division by which the number of groups L in each combination reduces one group. Then, we select among those combinations the $(L-1)$ -division which minimizes the within sum of squared errors, and let $\pi_{(L-1)}$ be the risk incurred by this selected division. Straightforwardly, we reduce the selected division one more group successively as long as its risk is strictly lesser than the risk of the previous division, and if not; the process stops. The final chosen division, that has $\hat{k}_0$ groups, is the desired solution because it has the smallest risk among the previous selected divisions and for this reason; we call it the most significant periods-solution.

To assess the proposed procedure when the data is free of outliers, the simulation achieves the objective quantities $\Pr(\hat{k}_0 = k_0)$ and *ANOMG* as seen in Table 4.1.

Table 4.1 Free outliers data & Most significant periods-solution.

k_0	3	4	5	6	7	
$\Pr(\hat{k}_0 = k_0)$	98.6%	99.8%	100%	99.6%	97.2%	Ave= 99.04%
ANOMG	0.01	0.02	0.02	0.02	0.03	Ave=0.02

The overall quantities 99.04% and 0.02 indicate the high reliability of the method in pure data. The first quantity is even more reliable than the result 96.5% in Aboukalam(2005).

Unfortunately, outliers may desert the achieved reliable results. If we repeat the experiments using contaminated data, the inequalities $M \geq k_0$ and $L \geq k_0$ were always true and the average of the differences $(M - k_0)$ and $(L - k_0)$ become 16.04 and 3.6 respectively, which indicate the low rapidity of the stopping rule. Table 4.2 represents the other objective quantities.

Table 4.2 Contaminated data & Most significant periods-solution

k_0	3	4	5	6	7	
$\Pr(\hat{k}_0 = k_0)$	58.6%	56.2%	51.4%	49.6%	49.2%	Ave=53%
ANOMG	1.24	2.25	2.65	3.42	7.43	Ave=3.4

It is remarkable that outliers decreased the overall percentage $\Pr(\hat{k}_0 = k_0)$ from 99.04% to 53%, and increased the error *ANOMG* from 0.02 to 3.4. So, the underlying technique is not robust because it does not resist outliers. Indeed, outliers should be detected and then be treated in order to avoid their bad effects on the results.

4.2. Detecting & Correcting Outliers

The simulation experiments in section 3 exhibits that any outlier has 80% chances to appear as a single group, or a group of size one, in the final *M*-division. In other words, the majority of the outliers in the data come out as single groups. To cure a questionable point, say *O*, which came out in a single group, we compare it with the means of its neighbors. Certainly, there would be either one neighbor or two neighbors. Suppose that; $\bar{X}$, v and S are the mean, volume and standard deviation of the nearer neighbor, respectively. We replace O by $\bar{X}$ if $v>1$ given that $|\bar{X} - O| \geq 5S$. But if $v=1$, the point O takes again $\bar{X}$ if there is only one neighbor, or in the other cases it takes the average of the two neighbors.

Once we complete the correction, we again find another optimal M-division. Straightforwardly, we do another correction and find a new M-division repeatedly until none of the single groups in the final M-division is subject to any correction. It is remarkable that none of the outliers could escape from the above correction scheme.

To assess the correcting scheme, let us repeat the simulation experiments again. Indeed, when the data is outliers free, the results are similar to Table 4.1. If the data is contaminated, then the correcting scheme turned the damaged results up, as in Table 4.3.

Table 4.3 Contaminated data & Most significant periods-solution & Correcting scheme

k_0	3	4	5	6	7	
$\Pr(\hat{k}_0 = k_0)$	95.4%	95.6%	96.4%	94.4%	90.6%	Ave=94.6%
ANOMG	0.21	0.32	0.44	0.56	0.70	Ave=0.45

The correcting scheme increased the probability $\Pr(\hat{k}_0 = k_0)$ from 53 % to 94.6%. This quantity is also more reliable than the result 91.6% achieved in Aboukalam(2005).

5. EXAMPLE

Table 5.1 shows the U.S combat deaths in Vietnam War during 72 months of the years 1966-1971.

Table 5.1[2] . U.S combat deaths in Vietnam War during 72 months of the years 1966-1971

	1966	1967	1968	1969	1970	1971
JAN	282	520	1202	795	343	140
FEB	435	662	2124	1073	386	221
MAR	507	944	1543	1316	449	272
APR	316	710	1410	847	526	226
MAY	464	1233	2169	1209	754	138
JUN	570	830	1146	1100	418	108
JUL	435	781	813	638	332	69
AUG	396	535	1080	795	319	67
SEP	419	775	1053	477	219	78
OCT	340	733	600	377	170	29
NOV	475	381	703	446	167	19
DEC	432	774	749	341	130	17

This is an ideal data that might be analyzed to detect the different periods of U.S involvement in Vietnam War. The data consists of some outliers. The stopping rule of the technique attained an M-division with $M = 33$ groups. Figure 5.1 and Table 5.2 exhibit this M-division in graphical and numerical ways.

[2] From Unclassified Statistics on Southeast Asia (1972), Department of Defense, OASD (Comptroller), Directorate for Information Operation

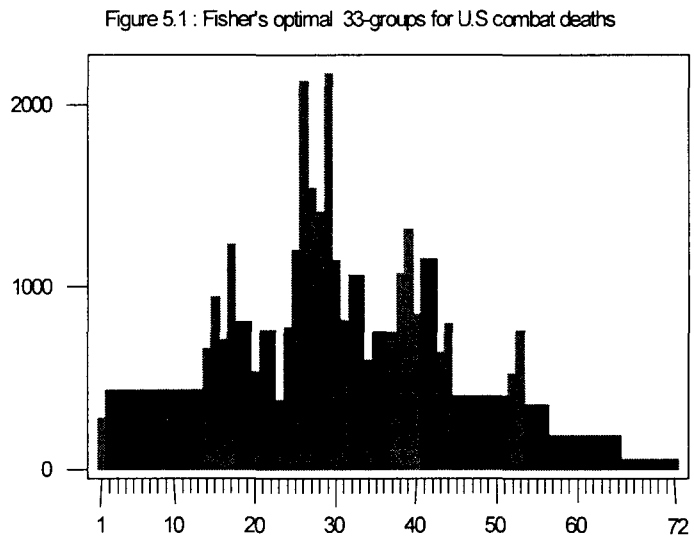

Figure 5.1 Fisher's optimal 33 groups for U.S. combat deaths

Indeed, any group whose size is one observation is questionable to be an outlier. If the technique does not correct these points, the initial free dwarfish L–division will have $L = 8$ groups, and the non robust most significant solution will have $\hat{k}_0 = 7$ periods, as seen in Table 5.3. Or else, if the questionable points should be corrected, the correcting scheme will perform 3 correcting circles to attain an M-division with M=12, see Table 5.4. Moreover, the technique will initiate a free dwarfish L -division with $L = 8$ groups and will reach to a robust most significant solution with $\hat{k}_0 = 7$ periods, see Table 5.5.

Table 5.2 Fisher's optimal division of 33- groups for the U.S combat deaths

Order	Group	Size	Mean of Deaths
1	1-1	1	282
2	2-13	12	437.2
3	14-14	1	662
4	15-15	1	944
5	16-16	1	710
6	17-17	1	1233

7	18-19	2	805.5
8	20-20	1	535
9	21-22	2	754
10	23-23	1	381
11	24-24	1	774
12	25-25	1	1202
13	26-26	1	2124
14	27-27	1	1543
15	28-28	1	1410
16	29-29	1	2169
17	30-30	1	1146
18	31-31	1	813
19	32-33	2	1066
20	34-34	1	600
21	35-37	3	749
22	38-38	1	1073
23	39-39	1	1316
24	40-40	1	847
25	41-42	2	1154.5
26	43-43	1	638
27	44-44	1	795
28	45-51	7	402.7
29	52-52	1	526
30	53-53	1	754
31	54-56	3	356.3
32	57-65	9	187
33	66-72	7	55.3

Table 5.3 Most significant periods-solution & without correcting scheme

Order	Periods	Size	Mean of Deaths
1	1 – 13	13	425.2
2	14 – 25	12	796.7
3	26 – 30	5	1678.4
4	31 – 37	7	827.6
5	38 – 42	5	1109.0
6	43 – 56	14	471.5
7	57 – 72	16	129.4

Table 5.4 Fisher's optimal division of 12- groups attained after 3 correcting circles.

Order	Group	Size	Mean of Corrected Deaths
1	1-1	1	282.0
2	2-13	12	437.2
3	14-24	11	759.1
4	25-26	2	1410.8
5	27-29	3	1758.1
6	30-33	4	1092.9
7	34-37	4	711.8
8	38-42	5	981.0
9	43-43	1	799.3
10	44-56	13	413.5
11	57-65	9	187.0
12	66-72	7	55.3

Table 5.5 Most significant periods-solution & with correcting scheme

Order	Periods	Size	Mean of Corrected Deaths
1	1 – 13	13	425.3
2	14 - 24	11	759.1
3	25 - 29	5	1619.2
4	30 - 43	14	923.1
5	44 - 56	13	413.5
6	57 - 65	9	187.0
7	66 - 72	7	55.3

The robust solution in Table 5.5 re-organizes some periods and observations in Table 5.3 in a better manner. In general, Period 4 in Table 5.5 was split in Table 5.3 into Periods 4 and 5. Moreover, Periods 6 and 7 in Table 5.5 were combined in Table 5.3 in Period 7. Indeed, a simple glance of Figure 5.1 shows that Robust decision in Table 5.5 is the best. Precisely in Table 5.5; Period 4 should not be split, and Periods 6 and 7 should not be combined. Eventually, Table 5.6 is Robust decision of Aboukalam(2005) that reached to 6 Periods. Unfortunately, the decision could not avoid the second fault. As a result, the underlying Robust most significant solution is the supreme.

Table 5.6 t-division solution in Aboukalam and with correcting scheme

Order	Periods	Size	Mean of Corrected Deaths
1	1 – 13	13	425.3
2	14 - 24	11	759.1
3	25 - 29	5	1619.2
4	30 - 43	14	923.1
5	44 - 56	13	413.5
6	57 - 72	16	129.4

ACKNOWLEDGMENT

This research was supported by the Research Center Project No (Stat/2006/07) in King Saud University.

REFERENCES

Aboukalam, M. A. F.(2005). Phases of Developments in Time ordered data with outliers treatment. JSTA., **4**, 272-280.

Bellman, R. E. and Dreyfus, S. E.(1962). *Applied Dynamic Programming*, Princeton University Press.

Fisher, W. D. (1958). On Grouping for Maximum Homogeneity. *Journal of The American Statistical Association*, **53**, 789-98.

Hartigan, J. A. (1975). *Clustering Algorithms*. Wiley, New York.