Outlier Detection Method using Covratio Statistics for Wind Direction Data with the Unreplicated Linear Functional Relationship Model and Equal Error Concentration Parameters

N. A. MOKHTAR[a], Y. Z. ZUBAIRI[b], A. G. HUSSIN[c] & A. A. GHAPOR[d]

[a,c] Faculty of Defence Sciences and Technology,

National Defence University of Malaysia,

Kem Sungai Besi, 57000 Kuala Lumpur, Malaysia.

[b]Centre for Foundation Studies in Science,

University of Malaya, 50603 Kuala Lumpur, Malaysia.

[d]Faculty of Economic and Administration,

University of Malaya, 50603 Kuala Lumpur, Malaysia.

PRESENTATION OUTLINE

- Abstract
- Introduction
- Literature Review
- Maximum Likelihood Estimation of Parameters
- Determining the Cut-off Point
- Power of Performance in Outlier Detection
- Application on Real Wind Direction Data
- Conclusion
- Reference

ABSTRACT

Wind direction data is importantly used in meteorology. The data is angular and to be more specific, it is circular. The relationship between the wind data may be studied using the unreplicated linear functional relationship model. However, there may happen to have outliers in the data. Therefore this paper is discussing about an outlier detection method for the unreplicated linear functional relationship model (LFRM) of circular variables with equal error concentration parameters. The covariance matrix is derived with some correction factor applied to the maximum likelihood estimation and the covratio statistics of the model is obtained. The cut-off point is developed based on the 5% upper percentile of the covratio statistics. The simulation study shows that the power of performance of outlier detection gets better as the degree of contamination gets bigger. The applicability of the proposed method is illustrated by using the wind direction data collected from the Holderness Coastline at the Humberside Coast in North Sea, United Kingdom.

INTRODUCTION

- **Wind Direction Data**
- The proposed method is applied to a real wind direction data with n = 129 obtained from Holderness Coastline, Humberside Coast, United Kingdom.

- Variable x is the data of the wind direction measured by HF radar system and developed by UK Rutherford and Appleton Laboratories, using the pulse radar and operates at frequency of 24.2-27 MHz.

- Meanwhile for the variable y, the data was measured by anchored wave buoy.

- The data is circular and the unit is Radian.

INTRODUCTION

➢ Circular statistics

- Data are measured in the range $[0°, 360°)$ or $[0, 2\pi)$ radians.
- Formal analysis cannot be done to circular data with usual statistical technique due to the wrap-around nature of a circle.
- The von Mises distribution is said to be the most useful distribution for circular data ***(Mardia and Jupp, 2000).***
- The p.d.f. of von Mises distribution: $$g(\theta;\mu,\kappa)=\frac{1}{2\pi I_0(\kappa)}e^{\kappa\cos(\theta-\mu)} \quad (1)$$

where $I_0(\kappa)=\frac{1}{2\pi}\int_0^{2\pi} e^{\kappa\cos\theta}d\theta;\ \ 0\le x<2\pi,\ 0\le\mu<2\pi,\ \kappa>0.$

LITERATURE REVIEW

- Examples of circular data:

  - Kendall (1974) studied the bird navigation ***(Best and Fisher, 1979).***
  - geologists studying the direction of the earth's magnetic pole ***(Batschelet, 1981).***
  - directions of the stars ***(Fisher,1993)***
  - In medical application, Jammaladaka et al. (1986) discussed about the angle of knee flexion to assess the recovery of ortheopaedic patients ***(Jammaladaka and Sengupta, 2001).***

LITERATURE REVIEW

➢ Unreplicated Linear Functional Relationship Model for Circular Data

- $\mathrm{Y} = \alpha + \beta X \ (mod\ 2\pi)$ (2)
- **Caires & Wyatt** : β=1, the model becomes symmetry
- $\mathrm{Y} = \alpha + \beta X \ (mod\ 2\pi)$ and $X = \alpha^* + \beta^* Y (mod\ 2\pi)$
- α is the rotation parameter
- x_i and y_i are subject to random errors δ_i and ε_i, respectively
- $x_i = X_i + \delta_i$ ⟶ error: $\delta_i \sim VM(0, \kappa)$
- $y_i = Y_i + \varepsilon_i$ ⟶ error: $\varepsilon_i \sim VM(0, v)$

LITERATURE REVIEW

- Outlier detection is the process of detecting the data object which is inconsistent or grossly different from the remaining set of data ***(Duan et al. ,2009).***
- **Outlier** : the one that appears to deviate markedly from other members of the sample in which it occurs ***(Grubbs, 1969).***
- If we remove the outliers, data will lead us to a different model.
- However, the distinction between these kinds of points is not always obvious ***(Rahman et al., 2012).***

MAXIMUM LIKELIHOOD ESTIMATION OF PARAMETERS

- There is an assumption of $\kappa = \nu$, thus $\lambda = \frac{\nu}{\kappa}$.

- The log-likelihood function of the von Mises distribution becomes

$$\log L(\alpha, \kappa, X; x, y) = -2n \log 2\pi - 2n \log I_0(\kappa) + \kappa \sum_{i=1}^{n} \cos(x_i - X_i) + \kappa \sum_{i=1}^{n} \cos(y_i - \alpha - X_i)$$

3

- $$\hat{\alpha} = \begin{cases} \tan^{-1}\left\{\frac{S}{C}\right\} & \text{when } S > 0, C > 0 \\ \tan^{-1}\left\{\frac{S}{C}\right\} + \pi & \text{when } C < 0 \\ \tan^{-1}\left\{\frac{S}{C}\right\} + 2\pi & \text{when } S < 0, C < 0 \end{cases}$$

4

where $S = \sum_{i=1}^{n} \sin(y_i - \hat{X}_i)$ and $C = \sum_{i=1}^{n} \cos(y_i - \hat{X}_i)$

MAXIMUM LIKELIHOOD ESTIMATION OF PARAMETERS

- $$\hat{\kappa} = A^{-1}\left(\frac{1}{n}\left\{\sum_{i=1}^{n}\cos\left(x_i - \hat{X}_i\right) + \sum_{i=1}^{n}\cos\left(y_i - \alpha - \hat{X}_i\right)\right\}\right) \quad (5)$$

- Caires and Wyatt (2003) noted that, in circular case, the estimation of a concentration parameter (whose inverse is equivalent of the variance for linear data) needs to be corrected by dividing it by 2. It has been proposed that $\tilde{\kappa} = \frac{\hat{\kappa}}{2}$ gives a better approximation to the value of κ.

- $$\hat{X}_{i1} \approx \hat{X}_{10} + \frac{\sin\left(x_i - \hat{X}_{i0}\right) + \sin\left(y_i - \hat{\alpha} - \hat{X}_{i0}\right)}{\cos\left(x_i - \hat{X}_{i0}\right) + \cos\left(y_i - \hat{\alpha} - \hat{X}_{i0}\right)} \quad (6)$$

COVARIANCE MATRIX OF PARAMETERS

- In 2015, Mokhtar et al. studied the unreplicated LFRM with the assumption of equal error concentration parameter and the covariance matrix of the parameter in the model is

$$\operatorname{cov}\begin{bmatrix}\hat{\alpha}\\ \tilde{\kappa}\end{bmatrix} = \begin{bmatrix}\dfrac{1}{2nA'(\tilde{\kappa})} & 0\\ 0 & \dfrac{2}{n\tilde{\kappa}A(\tilde{\kappa})}\end{bmatrix} \quad (7)$$

- Therefore, the determinant of the covariance matrix for this model is

$$|COV| = \frac{1}{n^2\tilde{\kappa}A(\tilde{\kappa})} \quad (8)$$

DETECTING OUTLIER USING *COVRATIO* STATISTICS

- The statistic is used to measure the effect of removing the observation based on the determinantal ratio given by:

$$COVRATIO_{(-i)} = \frac{|COV|}{|COV_{(-i)}|} \qquad (9)$$

where $|COV|$ is the determinant of covariance matrix for the full set and $|COV_{(-i)}|$ is the determinant of covariance matrix for the reduced data set by excluding the *i-th* row.

DETERMINING THE CUT-OFF POINT

- **Step 1**: Generate the values of X variable from the von Mises distribution of VM (2,3) and in the size of n = 20, 30, 50, 70, 100, 130 and 150. Find Y according to the generated X .
- **Step 2**: The variables X and Y are considered with generated random error terms of $\delta_i \sim VM(0,\kappa)$ and $\varepsilon_i \sim VM(0,\nu)$, respectively where $\kappa = \nu$.
- **Step 3**: The variables are fitted to the Unreplicated LFRM. The parameter estimates and the covariance of the parameters are calculated using the parameter estimation method in Mokhtar et al (2015).
- **Step 4**: Calculate the value of $|COV|$
- **Step 5**: Omit the i^{th} observation of the generated data, where i=1, 2, 3, …,n. Repeat steps 3-5 for all i to obtain $|COV_{(-i)}|$

DETERMINING THE CUT-OFF POINT

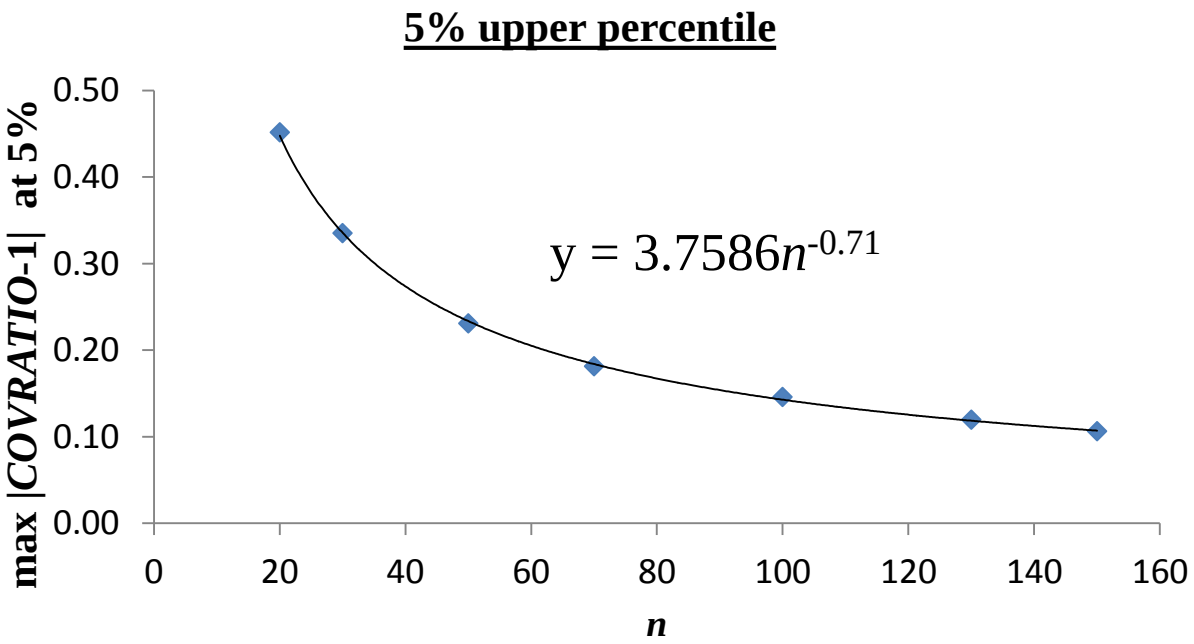
- **Step 6**: Calculate the value of $COVRATIO_{(-i)}$ and find $|COVRATIO_{(-i)} - 1|$ for all i.
- **Step 7**: Note the maximum value of $|COVRATIO_{(-i)} - 1|$.
- **Step 8:** The process is repeated for 500 simulations and the 5% upper percentiles of the maximum $|COVRATIO_{(-i)} - 1|$ are obtained.

Table 1 The values of 5% upper percentile of $|COVRATIO_{(-i)} - 1|$

n	κ =3	κ =5	κ =10	κ =15
20	0.482631	0.456601	0.427134	0.440119
30	0.345093	0.341747	0.327810	0.325791
50	0.247297	0.245086	0.213416	0.217016
70	0.194456	0.204701	0.162582	0.163800
100	0.134861	0.201583	0.126035	0.119783
130	0.113857	0.161380	0.104580	0.097920
150	0.101447	0.147530	0.091868	0.083569

DETERMINING THE CUT-OFF POINT

- The mean of the values for each *n* are calculated and the power series formula is plotted.
- We consider to obtain 95% confidence level and thus the cut-off point is to be at 5% significant level.
- Thus, the cut-off point is $y = 3.7586n^{-0.71}$. 11



POWER OF PERFORMANCE IN OUTLIER DETECTION

- **Step 1**: The values of X variable are generated from the von Mises distribution of VM and in the size of n = 30, 70, 100 and 130 and one observation X_d^* is then contaminated with some levels of contamination where the level of the contamination are ω = 0.2, 0.4, 0.6, 0.8 and 1. The formula of contaminating the observation is as follows:

$$X_d^* = X_i + \omega\pi \pmod{2\pi} \quad (12)$$

- **Step 2**: Find Y according to the generated X. The variables X and Y are considered with generated random error terms of $\delta_i \sim VM(0,\kappa)$ and $\varepsilon_i \sim VM(0,\nu)$, respectively where $\kappa = \nu$.
- **Step 3**: The variables are fitted to the Unreplicated LFRM. The parameter estimates and the covariance of the parameters are calculated. **Step 4**: Calculate the value of $|COV|$.

POWER OF PERFORMANCE IN OUTLIER DETECTION

- **Step 5**: Omit the i^{th} observation of the generated data, where i=1, 2, 3, …,n. Repeat steps 3-5 for all i to obtain $|COV_{(-i)}|$.
- **Step 6**: Calculate the value of $COVRATIO_{(-i)}$ and find $|COVRATIO_{(-i)} - 1|$ for all i.
- **Step 7**: Note if the maximum value of $|COVRATIO_{(-i)} - 1|$ exceeds the cut-off point and it is remarked as detecting the contaminated observation.
- **Step 8:** The percentage of correct detection of the outlier is calculated as the power of performance.

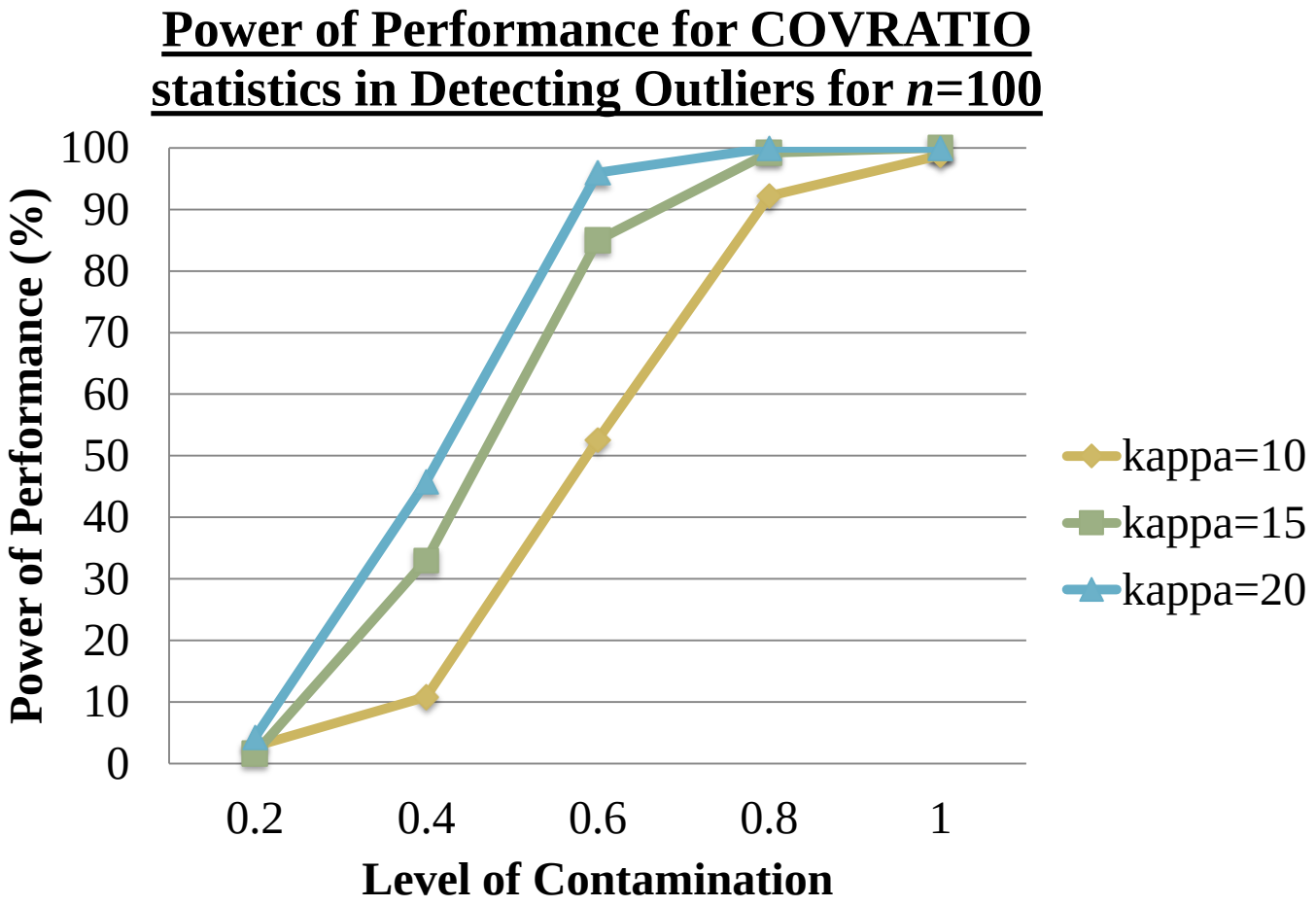
POWER OF PERFORMANCE IN OUTLIER DETECTION

- Table 2 the power of performance

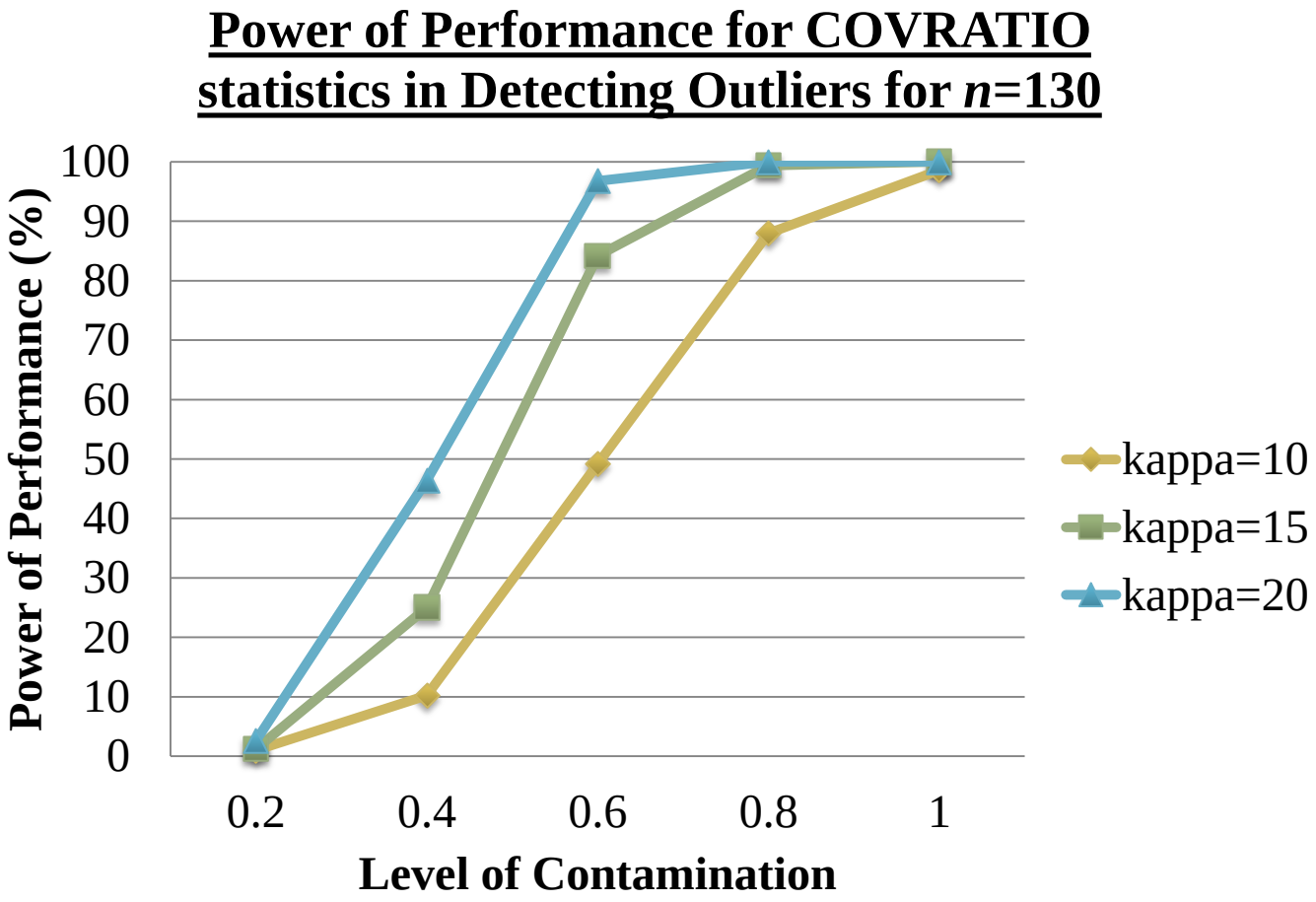
ω	n	κ =10	κ =15	κ =20
0.2	30	4.60	6.40	9.20
	70	1.80	3.40	6.00
	100	2.80	1.60	4.20
	130	1.00	1.20	2.60
0.4	30	21.60	41.80	60.60
	70	14.20	31.20	54.60
	100	10.80	33.00	45.80
	130	10.20	25.00	46.40
0.6	30	66.20	85.00	99.00
	70	64.00	86.40	97.40
	100	52.60	85.00	96.00
	130	49.20	84.20	96.80
0.8	30	94.20	99.80	100.00
	70	93.60	99.60	100.00
	100	92.20	99.20	100.00
	130	88.00	99.40	100.00
1	30	99.20	100.00	100.00
	70	99.40	100.00	100.00
	100	98.80	100.00	100.00
	130	98.60	100.00	100.00

- The power of performance increases as the concentration parameter and the level of contamination increase.
- The highest concentration parameter and highest level of contamination give the highest power of performance where the value is 100.

POWER OF PERFORMANCE IN OUTLIER DETECTION



POWER OF PERFORMANCE IN OUTLIER DETECTION



APPLICATION ON REAL WIND DIRECTION DATA

- The proposed method is applied to a real wind direction data set with n = 129 obtained from Holderness Coastline, Humberside Coast, United Kingdom.
- Previous researchers of circular statistics such as Abuzaid et al. (2008), Hussin et al. (2010) and Shamsudheen (2014) have used it and have established that observations 38 and 111 as outliers .
- Mokhtar et al. (2015) noted that this data is suitable to be described by the Caires and Wyatt model.
- Through the proposed method, observations 38 and 111 are detected as the outliers of the data set. The value of for these two observations exceed the cut-off point $y = 3.7586n^{-0.71}$.

APPLICATION ON REAL WIND DIRECTION DATA

- Cut-off point $y = 3.7586n^{-0.71} = 3.7586(128)^{-0.71} = 0.119924$

obs	$\lvert COVRATIO_{(-i)} - 1 \rvert$
1	0.011148
2	0.013208
3	0.007899
4	0.008425
5	0.009881
6	0.007789
7	0.007805
8	0.039076
9	0.007822
10	0.007998
11	0.009947
12	0.008832
13	0.007864
14	0.009639
15	0.019374
16	0.039371
17	0.013319
18	0.028767
19	0.010495
20	0.007781
21	0.011246
22	0.008583

obs	$\lvert COVRATIO_{(-i)} - 1 \rvert$
23	0.008819
24	0.009238
25	0.009335
26	0.008017
27	0.033682
28	0.019152
29	0.008023
30	0.009436
31	0.008524
32	0.008964
33	0.010548
34	0.007778
35	0.007778
36	0.008261
37	0.008105
38	0.326310
39	0.008739
40	0.014525
41	0.008128
42	0.036930
43	0.001128
44	0.008004

obs	$\lvert COVRATIO_{(-i)} - 1 \rvert$
45	0.008342
46	0.007778
47	0.008149
48	0.027849
49	0.017451
50	0.015385
51	0.007783
52	0.007788
53	0.009002
54	0.007802
55	0.007827
56	0.007802
57	0.038041
58	0.007791
59	0.007973
60	0.013301
61	0.009429
62	0.008414
63	0.008805
64	0.008290
65	0.011823
66	0.037421

obs	$\lvert COVRATIO_{(-i)} - 1 \rvert$
67	0.010282
68	0.025302
69	0.037346
70	0.043068
71	0.007877
72	0.016204
73	0.015715
74	0.008199
75	0.009254
76	0.008964
77	0.010166
78	0.008771
79	0.008656
80	0.037346
81	0.013711
82	0.009402
83	0.007998
84	0.007878
85	0.009986
86	0.007797
87	0.008860
88	0.009463

obs	$\lvert COVRATIO_{(-i)} - 1 \rvert$
89	0.008771
90	0.009958
91	0.014036
92	0.009629
93	0.007775
94	0.012077
95	0.029329
96	0.015826
97	0.020960
98	0.033493
99	0.035981
100	0.023925
101	0.008007
102	0.008125
103	0.008779
104	0.008950
105	0.007973
106	0.007782
107	0.008453
108	0.008185
109	0.037488
110	0.008112

obs.	$\lvert COVRATIO_{(-i)} - 1 \rvert$
111	0.137758
112	0.007883
113	0.008220
114	0.007814
115	0.008706
116	0.007824
117	0.011148
118	0.007985
119	0.009908
120	0.011370
121	0.008832
122	0.007773
123	0.031913
124	0.008234
125	0.008752
126	0.007860
127	0.009832
128	0.011028
129	0.010249

CONCLUSION

- The covariance matrix is derived with some correction factor applied to the maximum likelihood estimation to obtain the covratio statistics of the model.

- 95% confidence level is considered and thus the cut-off point developed is to be at 5% significant level.

- The pattern in the power of performance in the simulation study shows that this method is adequate to detect the outlier exists in a circular data.

REFERENCE

- Abuzaid A. H., Hussin A. G. and Mohamed I. B. 2008. Identifying single outlier in linear circular regression model based on circular distance. Journal of Applied Probability & Statistics: 3(1), 107-117
- Abuzaid, A., Mohamed I, Hussin, A. G. and Rambli, A. (2011) "Covratio Statistics for Simple Circular Regression Model", *Chiang Mai Journal of Science*, 38 (3), 321-330
- Belsley, D. A., Kuh, E and Welsch, R. E. (1980). Regression Diagnostics: Identifying Influential Data and Sources of Collinearity, Wiley, New York.
- Caires, S. and Wyatt, L. R. (2003) "A Linear Functional Relationship Model for Circular Data with an Application to the Assessment of Ocean Wave Measurement", American Statistical Association and the Internal Biometric Society Journal of Agricultural, Biological and Environmental Statistics, 8 (2),153-169.
- Fisher, N. I. (1993). Statistical Analysis of Circular Data, Cambridge University Press.
- Ghapor, A. A, Zubairi, Y. Z, Mamun, A. S. M. A. and Imon, A. H. M. R. (2014) "On Detecting Outlier in Simple Linear Functional Relationship Model using Covratio Statistic", *Pakistan Journal of Statistics*, 30(1), 129-142
- Hassan, S. F, Hussin, A. G. and Zubairi, Y. Z. (2010) "Estimation of Functional Relationship Model for Circular Variables and Its Application in Measurement Problem", Chiang Mai Journal of Science, 37(2), 195-205.

REFERENCE

- Hussin A.G, Abuzaid A, Zulkifli F. and Mohamed I. (2010) Asymptotic Covariance and Detection of Influential Observation in a Linear Functional Relationship Model for Circular Data with Application to the Measurements of Wind Directions, Science Asia: 36, 249-253.
- Ibrahim S, Rambli A, Hussin A G and Mohamed I, (2013) "Outlier Detection in a Circular Regression Model Using Covratio Statistic", *Communication in Statistics-Simulation and Computation*, 42, 2272-2280
- Mardia, K. V. and Jupp, P. E. (2000). Directional Statistics, John Wiley & Sons.
- Mokhtar, N. A., Zubairi, Y. Z. and Hussin, A. G. (2015) "A Simple Linear Functional Relationship Model for Circular Variables and its Application" *Proceedings of the 9th International Conference on Renewable Energy Sources* (RES '15), Kuala Lumpur, Malaysia. (Page 57-63).
- Rambli, A., Abuzaid, A. H. M. , Mohamed, I. B. and Hussin, A. G. (2016) "Procedure for Detecting Outliers in a Circular Regression Model", *PLOS ONE*, 11(4) e0153074.
- Shamsudheen M. I. (2014) Bootstrapping and Outlier Detection Problems in Linear Functional Relationship Model for Circular Data, MSc. Thesis, Universiti Pertahanan Nasional Malaysia.

THANK YOU