UNIVERSITY OF DUBLIN

XST30071

TRINITY COLLEGE

FACULTY OF ENGINEERING, MATHEMATICS AND SCIENCE

School of Computer Science and Statistics

JS MSISS, JS-SS MATHS & TSM **Trinity Term 2012**

Multivariate Linear Analysis and Applied Forecasting (ST3007)

Thursday, May 10, 2012 **RDS-MAIN** **9.30 — 12.30**

Dr. Rozenn Dahyot & Dr. Brett Houlding

Attempt two questions out of three in each section A and B

All questions carry equal marks

Non-programmable calculators are permitted for this examination—please indicate the make and model of your calculator on each answer book used.

You may not start this examination until you are instructed to do so by the Invigilator.

Section A - Multivariate Linear Analysis

1) The location, depth and magnitude of 50 significant seismic events that have occurred near Fiji since 1964 are recorded. Hierarchical and iterative cluster analyses were performed to establish if there existed groups within the data, with the outputs provided in Appendices A1 and A2 at the end of this question.

Variable	**Description**
lat	Latitude of event
long	Longitude of event.
depth	Depth of event (km).
mag	Richter magnitude of event.

a) Within Hierarchical clustering, choices are available concerning the dissimilarity measure and the linkage function. Explain what effect both of these have, and list and provide the definition of three possibilities for both.

[6 marks]

b) Detail the *k*-means cluster algorithm.

[3 marks]

c) Explain what shape of clusters *k*-means is effective at finding and also why it is important to run the algorithm from multiple initializations.

[4 marks]

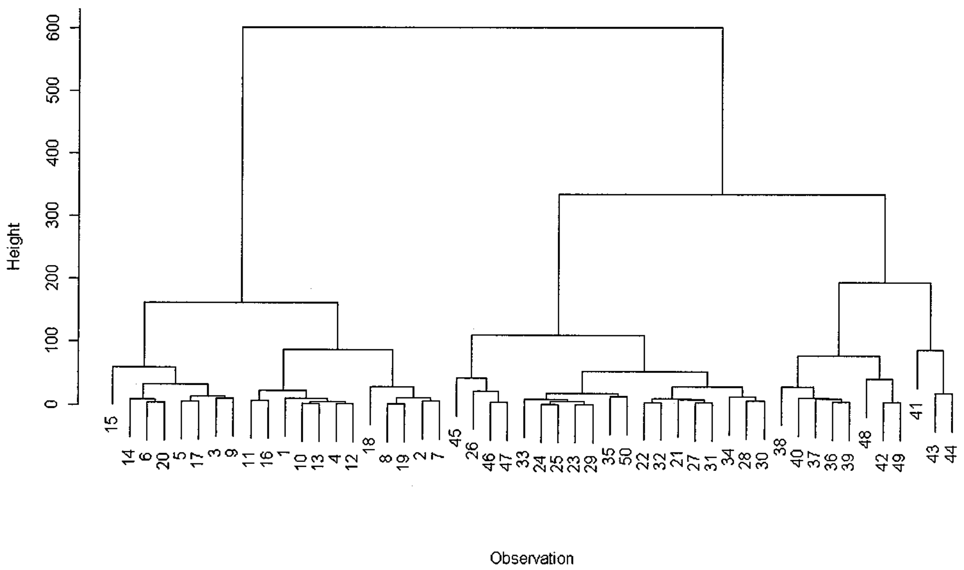
d) Describe what conclusions can be drawn from the cluster analyses in Appendices A1 and A2. In particular, what is the appropriate number of clusters for this data and what can be said about those clusters?

[8 marks]

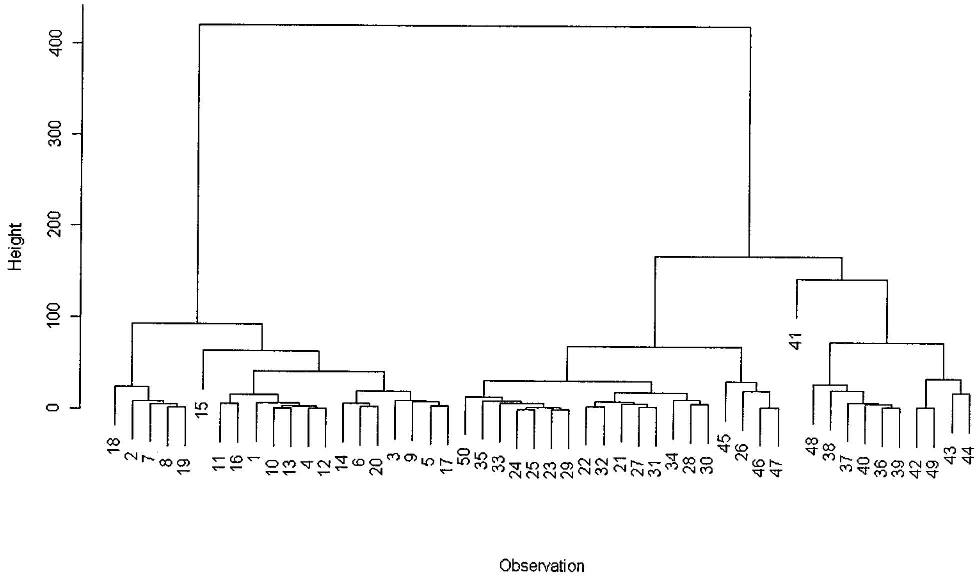
e) Briefly describe why the use of standardization and dimension reduction may be of benefit if applied to the data before a cluster analysis is performed.

[4 marks]

Appendix A.1



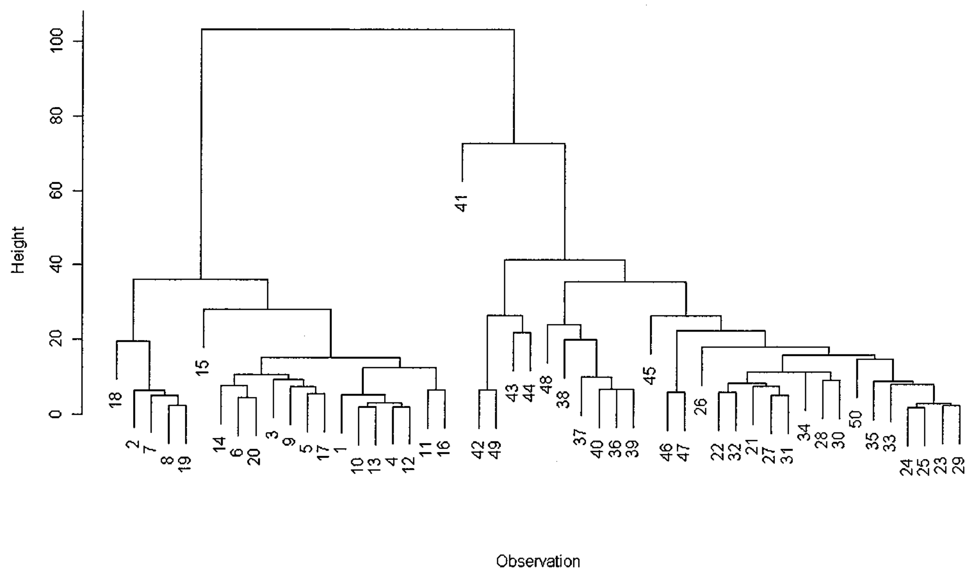
Dendrogram with Complete Linkage and Euclidean Distance



Dendrogram with Average Linkage and Euclidean Distance

Continued on the next page

Dendrogram with Single Linkage and Euclidean Distance



Appendix A.2

Number of clusters: 1

	Number of Obs.	Within Sum of Squares	Avg. Distance From Centroid
Cluster 1	50	2423669	204.97
Sum	50	2423669	

Cluster Centroids:

	Cluster 1	Total Data
lat	-21.47	-21.47
long	181.91	181.91
depth	306.2	306.2
mag	4.58	4.58

Distance Between Cluster Centroids:

	Cluster 1
Cluster 1	0.00

Continued on the next page

```
Number of clusters:     2

                 Number of Obs.        Within Sum of Squares  Avg Distance From Centroid
Cluster 1             21                     76121                  46.77
Cluster 2             29                     197092                 72.82
Sum                   50                     273213

Cluster Centroids:
         Cluster 1        Cluster 2        Total Data
lat      -20.45           -22.20           -21.47
long     181.20           182.43           181.91
depth    549.90           129.72           306.2
mag      4.50              4.64            4.58

Distance Between Cluster Centroids:

                 Cluster 1        Cluster 2
Cluster 1        0.00             420.19
Cluster 2        420.19           0.00

-----
Number of clusters:     3

                 Number of Obs.        Within Sum of Squares  Avg Distance From Centroid
Cluster 1             11                     35539                  48.11
Cluster 2             19                     21744                  28.31
Cluster 3             20                     44361                  37.99
Sum                   50                     101644

Cluster Centroids:
         Cluster 1        Cluster 2        Cluster 3        Total Data
 lat     -22.73           -21.70           -20.55           -21.47
long     181.31           183.07           181.14           181.91
depth    244.81           76.05            558.60           306.2
mag      4.59              4.68             4.49            4.58

Distance Between Cluster Centroids:

                 Cluster 1        Cluster 2        Cluster 3
Cluster 1        0.00             168.78           313.79
Cluster 2        168.78           0.00             482.55
Cluster 3        313.79           482.55           0.00
```

Continued on the next page

Number of clusters: 4

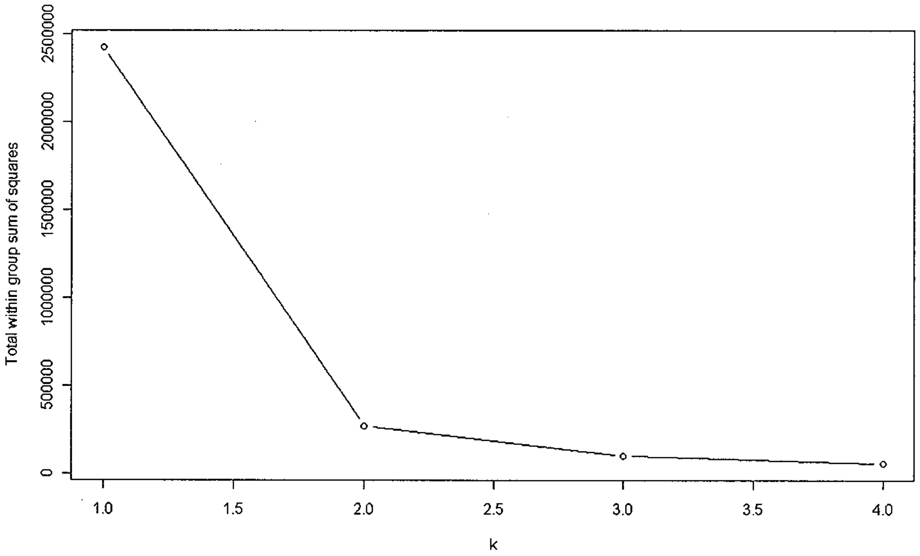
	Number of Obs.	Within Sum of Squares	Avg. Distance From Centroid
Cluster 1	6	11196	31.60
Cluster 2	10	16588	37.53
Cluster 3	15	11505	23.89
Cluster 4	19	21744	28.31
Sum	50	61034	

Cluster Centroids:

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Total Data
lat	-21.59	-23.15	-20.00	-21.70	-21.70
long	181.47	181.21	181.10	183.07	183.07
depth	470.12	231.70	581.80	76.05	76.06
mag	4.45	4.57	4.53	4.78	4.68

Distance Between Cluster Centroids:

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Cluster 1	0.00	238.47	111.65	394.12
Cluster 2	238.47	0.00	350.11	155.67
Cluster 3	111.65	76.05	0.00	505.75
Cluster 4	394.12	155.67	505.75	0.00



2) Six tests were performed by 112 individuals. The tests were described as:

• **general:**	a non-verbal measure of general intelligence
• **picture:**	a picture-completion test
• **blocks:**	block design
• **maze:**	mazes
• **reading:**	reading comprehension
• **vocab:**	vocabulary

Output from an application of Factor Analysis for this data is provided in Appendix B.

a) State and explain the Factor model for multivariate data $\mathbf{x}^T=(x_1, x_2, \ldots, x_m)$. That is to say, show how such multivariate data relates to the *common factors* and the *specific factors*, and indicate any assumptions about the expectation and covariance of these factor terms.

[5 marks]

b) In relation to the *variance* of a multivariate random variable, explain what is meant by the *communality* and *uniqueness* within a Factor Analysis, and show how these three terms are related.

[5 marks]

c) Explain what is meant by a *factor rotation*, why it is relevant within Factor Analysis, and state the objective of the *varimax* rotation.

[5 marks]

d) Explain what the output of Appendix B tells us about the test data.

[5 marks]

e) Describe two other dimension reduction techniques that could also have been applied to this data and how they differ from Factor Analysis.

[5 marks]

Appendix B

```
> Call: factanal(factors = 2, covmat = ability.cov, rotation = "varimax")

Uniquenesses:

general    picture    blocks    maze      reading    vocab
0.455      0.589      0.218     0.769     0.052      0.334

Loadings:
                  Factor1    Factor2
general           0.499      0.543
picture           0.156      0.622
blocks            0.206      0.860
maze              0.109      0.468
reading           0.956      0.182
vocab             0.785      0.225

                  Factor1    Factor2
SS loadings       1.858      1.724
Proportion Var    0.310      0.287
Cumulative Var    0.310      0.597
```

3) Data were recorded on 7,185 students.

Variable	Description
sexMale	1 if Male, 0 otherwise.
SES	A numeric score of socio-economic status.
MathAch	A numeric score of mathematics achievement.
Minority	1 if from a Minority Group, 0 otherwise.

Of interest is the relationship between minority status and the other recorded variables. The output from a logistic regression is given in Appendix C at the end of this question.

a) Making reference to the 'logit link function', explain why logistic regression, rather than linear regression, was used for the analysis of Appendix C.

[3 marks]

b) Explain what the output from Appendix C tells us about the relationship in this model between minority status and the *sexMale*, *SES* and *MathAch* scores.

[6 marks]

c) Give the formula for the probability that *Minority*=1 that results from the logistic model. Use this formula along with the output from Appendix C to predict the probability that a male with a *SES* value of 0.6 and a *MathAch* score of 9 will be of minority status.

[6 marks]

d) Making reference to the above example, describe the role and use of interactions in logistic regression.

[5 marks]

f) Describe two other classification techniques that could also have been applied to this data and how they differ from logistic regression.

[5 marks]

Appendix C

```
Call:

glm(formula = Minority ~ Sex + SES + MathAch, family = binomial(logit), data =
MathAchieve)

Coefficients:
             Estimate   Std. Error  z value    Pr(>|z|)
(Intercept)  -0.269981  0.061937    -4.359     1.31e-05 ***
SexMale      0.113598   0.056626    2.006      0.0448 *
SES          -0.625020  0.039104    -15.984    < 2e-16 ***
MathAch      -0.069133  0.004393    -15.736    < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4)

Section B - Applied Forecasting

4. (a) Give a definition for each of the following terms and explain their importance in the context of time series analysis:

   (i) RMSE and MAPE

   [4 marks]

   (ii) ACF and PACF

   [4 marks]

   (iii) ARIMA

   [4 marks]

   (b) Write the following equations without the backshift operator B:

   (i) $(1 - B^2)y_t = \epsilon_t$

   [2 marks]

   (ii) $(1 - \phi_1 B - \phi_2 B^2)(1 - B)y_t = (1 - \theta_1 B)(1 - B^{12})\epsilon_t$

   [3 marks]

   (c) Write the following equations with the backshift operator B, y_t and ϵ_t:

   (i) $y_t - y_{t-12} = \epsilon_t$

   [2 marks]

   (ii) $y_t - \phi_1\ y_{t-1} - \phi_2\ y_{t-12} = \theta_1 \epsilon_{t-1} + \epsilon_t$

   [3 marks]

   (d) What are the R commands used to

   (i) fit an ARIMA model?

   (ii) fit an exponential smoothing model?

   (iii) compute a forecast?

   (iv) plot the time series and the corresponding ACF and PACF plots?

   [3 marks]

   (25 marks)

5. (a) A time series denoted $\{y_t\}$ follows an ARIMA$(0,0,0)(0,0,1)_{12}$ model. What are the statistical hypotheses for this model? What is the mathematical expression of this model? (Use this equation for the following questions).

[4 marks]

(b) What is the expectation $\mathbb{E}[y_t]$ of the time series y_t?

[3 marks]

(c) Assuming now that $\mathbb{E}[y_t] = 0$, what is the variance of this time series y_t?

[2 marks]

(d) What is the covariance $\mathbb{C}\mathrm{ov}[y_t, y_{t-k}]$, $\forall k$?

[2 marks]

(e) What is the correlation $\mathbb{C}\mathrm{orr}[y_t, y_{t-k}]$, $\forall k$?

[2 marks]

(f) Comment on the shape of the ACF for this model.

[4 marks]

(g) What is the shape of the PACF for such model? Explain.

[2 marks]

(h) Assuming that we have collected n observations of this time series $\{y_t\}_{t=1,\cdots,n}$. What is the forecast $\hat{y}_{n+k}$ and its 95% confidence interval for $k = 1, \cdots, 12$?

[3 marks]

(i) What is the forecast $\hat{y}_{n+k}$ and its 95% confidence interval for $k = 13, \cdots, 24$?

[3 marks]

(25 marks)

6. A simple exponential smoothing model and an autoregressive model of order 1 were fitted to data concerning the number of strikes in the USA from 1951 to 1980. Details of the data analysis are shown in Figures 1, 2, 3 and 4. Answer all of the following questions with reference to the output where necessary.

   (a) Define the two models mathematically.

   [8 marks]

   (b) Are these two models appropriate for this data?

   [4 marks]

   (c) Which model do you think fits best? Why?

   [2 marks]

   (d) Explain how s^2 is computed in the AR(1) output.

   [4 marks]

   (e) Compute the predictions for the next 2 years based on each of the models.

   [4 marks]

   (f) Compute prediction intervals for the next 2 years based on the AR(1) model.

   [3 marks]

   (25 marks)

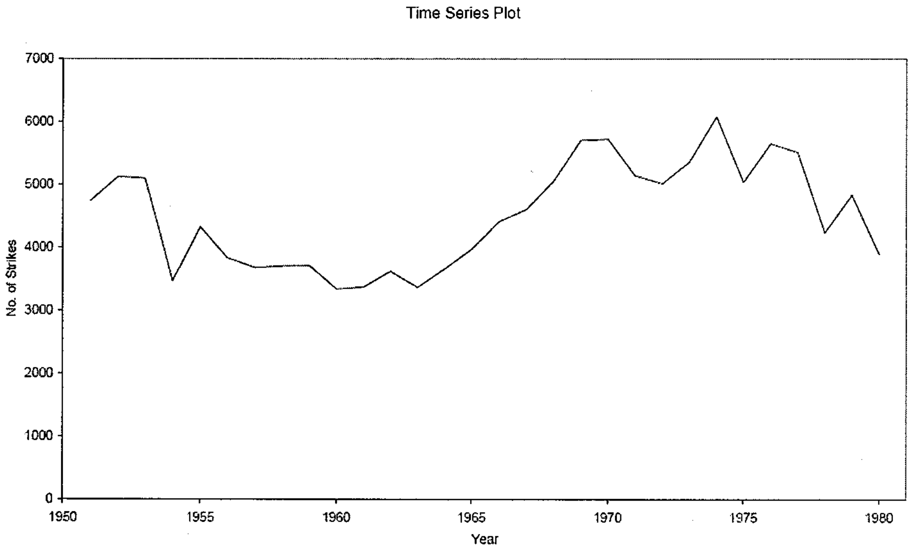
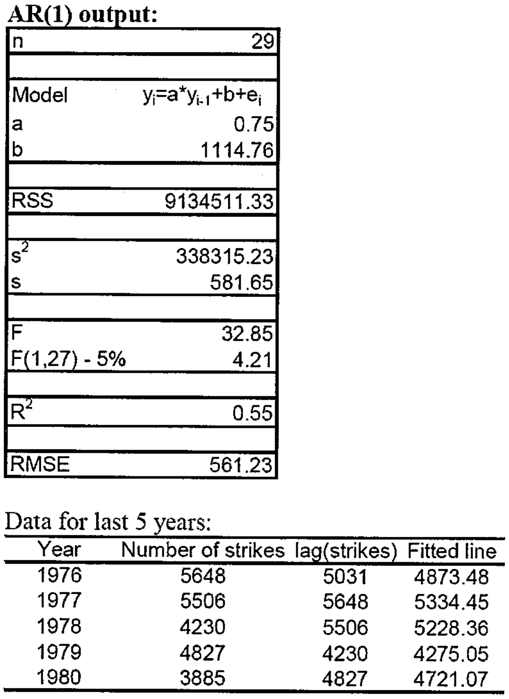


Figure 1: USA strikes



AR(1) output:

n	29
Model	$y_i=a*y_{i-1}+b+e_i$
a	0.75
b	1114.76
RSS	9134511.33
s^2	338315.23
s	581.65
F	32.85
F(1,27) - 5%	4.21
R^2	0.55
RMSE	561.23

Data for last 5 years:

Year	Number of strikes	lag(strikes)	Fitted line
1976	5648	5031	4873.48
1977	5506	5648	5334.45
1978	4230	5506	5228.36
1979	4827	4230	4275.05
1980	3885	4827	4721.07

Figure 2: USA strikes AR(1)

SES output:

t	30
alpha	0.68
RMSE	555.81
MAPE	0.10

Last 5 years:

Year	Number of strikes	Forecast for year t
1976	5648	5281.39
1977	5506	5531.59
1978	4230	5514.13
1979	4827	4637.75
1980	3885	4766.91

Figure 3: USA strikes SES

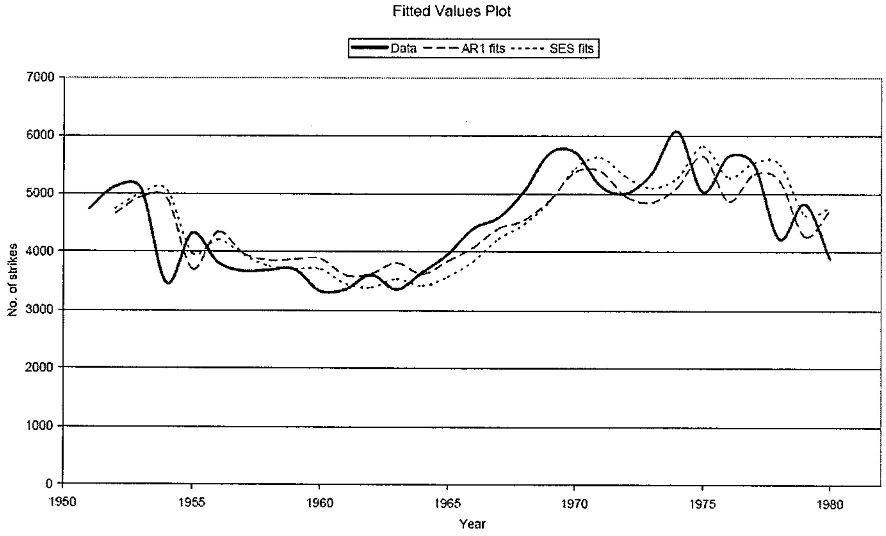


Figure 4: USA strikes (fitted values).