UNIVERSITY OF DUBLIN

XST3701

TRINITY COLLEGE

FACULTY OF ENGINEERING, MATHEMATICS AND SCIENCE

School of Computer Science and Statistics

JS-SS MSISS & Maths **Hilary Term 2009**

Multivariate Linear Analysis and Applied Forecasting (ST370)

Friday, 5 June, 2009 **GMB** **09:30 -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. Data were recorded on the educational transition of 474 Irish school children aged 11 in 1967.

Variable	Description
lvcert	Indicator variable with value 1 if Leaving Certificate taken (0 otherwise).
DVRT	Drumcondra Verbal Reasoning Test Score of child.
sex	Sex of the child (value 2 for female and 1 for male).
fathocc	A prestige score for the father's occupation.

Logistic regression was used to determine which factors were good predictors of whether a child would take a Leaving Certificate. Output from the logistic regression is given in Appendix A.1.

a) Provide a motivation for the use of logistic regression, rather than linear regression, for this data.

[3 marks]

b) Explain what the output from Appendix A.1 tells us about the relationship in this model between whether a child takes a Leaving Certificate and the other variables.

[6 marks]

c) Give the formula for the probability that *lvcert*=1 that results from the logistic model of Appendix A.1. Use this formula along with the output from Appendix A.1 to predict the probability that a female with *DVRT* score of 120 and a *fathocc* score of 30 will take a Leaving Certificate.

[6 marks]

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

[5 marks]

e) Briefly outline the difference between logistic regression and linear discriminant analysis as a classification tool.

[5 marks]

2. The location, depth and magnitude of 50 significant seismic events that have occurred near Fiji since 1964 are recorded.

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

Summary statistics are given in Appendix B.1.

Geo-physicists are interested in determining if there exist groups of seismic events that exhibit similar properties.

A Cluster Analysis was performed to establish if groups existed within the data. Outputs from three hierarchical cluster analyses are given in Appendix B.2, and from *k*-Means clustering in Appendix B.3.

a) Describe the hierarchical clustering methods, making reference to the term "linkage", and compare the output from the three hierarchical clustering methods.

[6 marks]

b) Provide a description of the *k*-Means clustering algorithm and describe what type of cluster structure *k*-Means clustering is effective at finding.

[6 marks]

c) Provide a description of the *k*-Means clustering output. In particular, describe what the numerical summaries tell us about the clusters for an appropriate value of *k*. Justify your choice of *k*.

[9 marks]

d) Give a brief account on the use of standardization and data reduction in conjunction with cluster analysis methods.

[4 marks]

3.

a) Briefly describe the objective of Multidimensional Scaling and the difference between its Metric and Non-Metric versions.

[4 marks]

b) Explain the meaning and role of the 'Stress' value for a Multidimensional Scaling Analysis and explain how this value is found. (You need only discuss standard Stress, and need not consider Sammon Stress or Kruskal Stress).

[4 marks]

c) Making reference to the Procrustes Sum of Squares, explain the role and application of a Procrustes Analysis within the context of Multidimensional Scaling.

[4 marks]

d) Contrast the similarities and differences between the approaches of Classical Metric Multidimensional Scaling and Principle Components.

[6 marks]

e) The age (days since 1968/12/31) and circumference (mm) for 35 Orange trees was recorded. The covariance matrix and output from two principal components analyses is given in Appendix C.1. Discuss the results of these analyses.

[7 marks]

Appendix A.1

Call:
glm(formula = lvcert ~ DVRT + sex + fathocc, family = binomial(logit))

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.1557	-0.9151	-0.4497	0.9175	2.2907

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-8.648086	0.984865	-8.781	< 2e-16 ***
DVRT	0.060585	0.008155	7.429	1.09e-13 ***
sex	0.516547	0.215021	2.402	0.0163 *
fathocc	0.039139	0.007492	5.224	1.75e-07 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

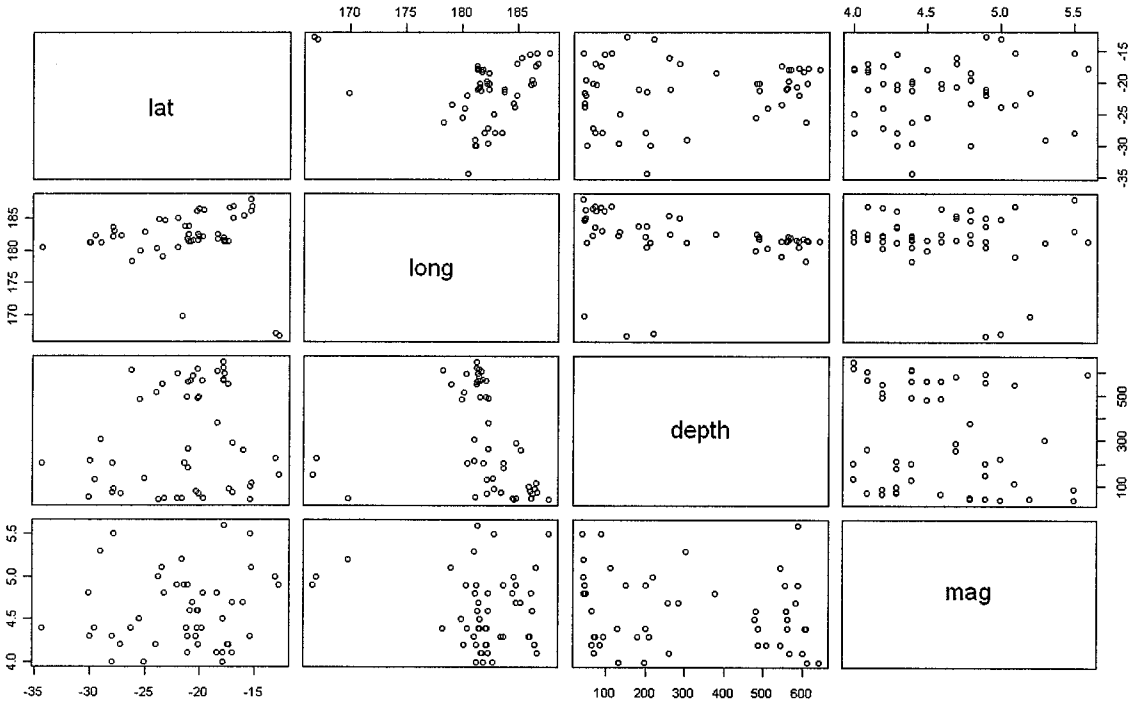
Null deviance:	651.39 on 473 degrees of freedom
Residual deviance:	530.63 on 470 degrees of freedom
AIC:	538.63

Number of Fisher Scoring iterations: 4

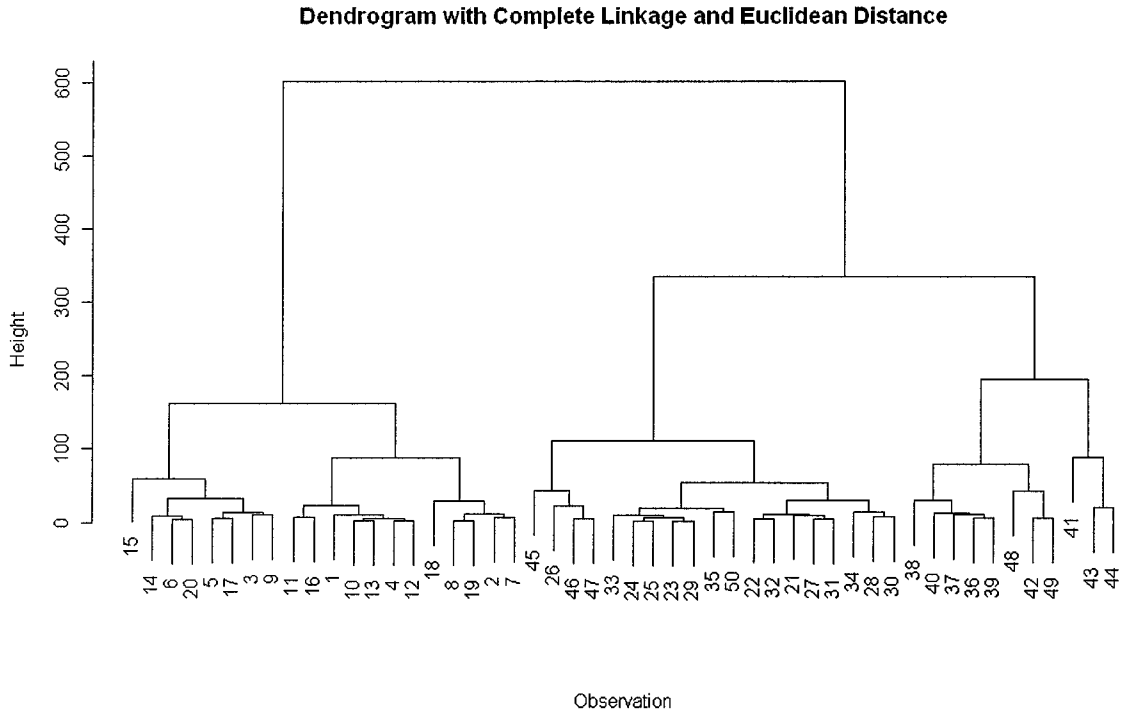
Appendix B.1

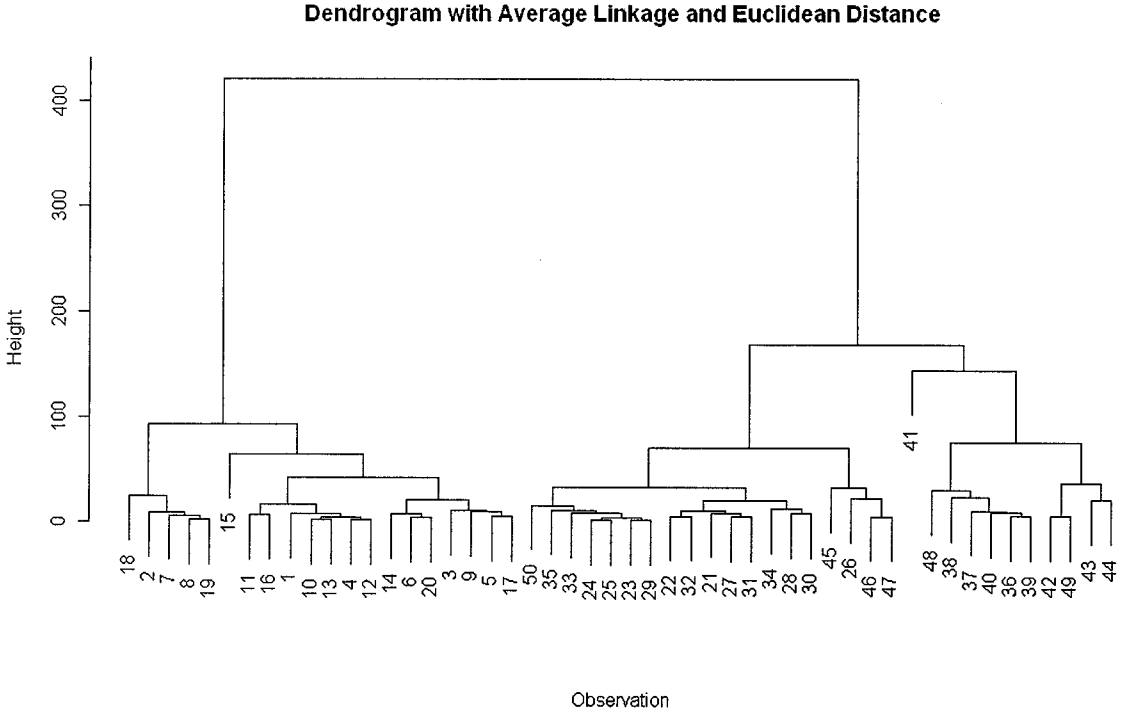
Variable	Mean	Standard Deviation
lat	-21.46	4.77
long	181.9	4.19
depth	306.2	222.3
mag	4.58	0.43

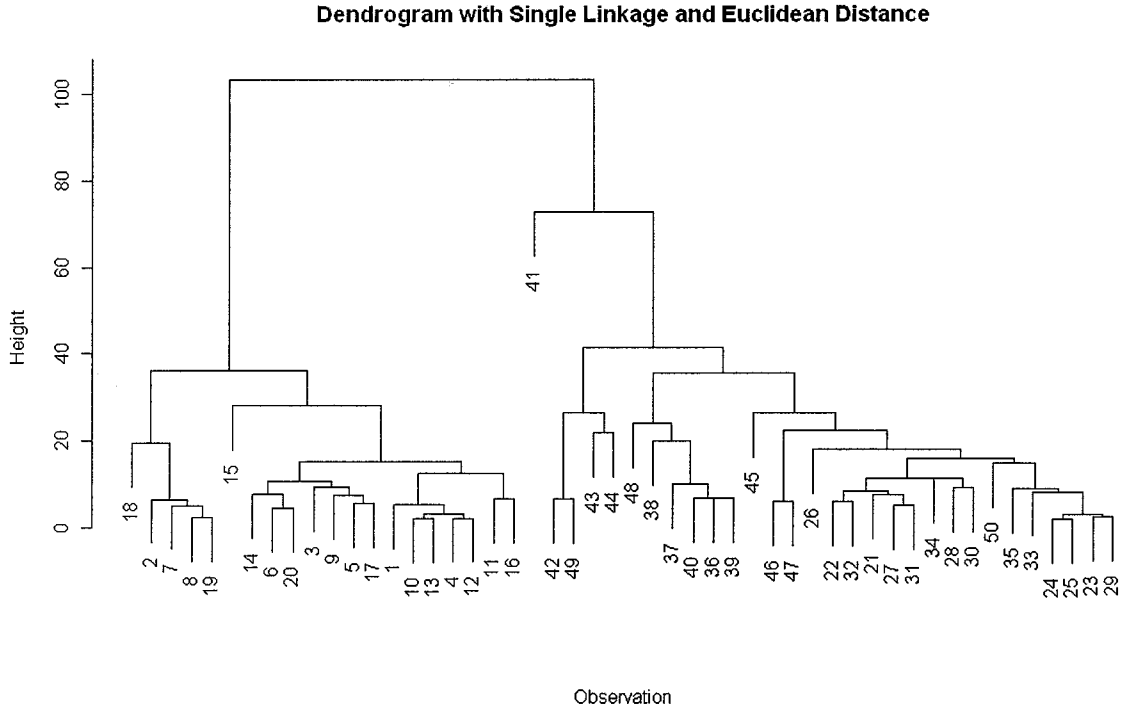
Matrix plot of lat, long, depth and mag



Appendix B.2


Dendrogram with Complete Linkage and Euclidean Distance
Height
0
100
200
300
400
500
600
Observation


Dendrogram with Average Linkage and Euclidean Distance
Height
0
100
200
300
400
Observation


Dendrogram with Single Linkage and Euclidean Distance
Height
0
20
40
60
80
100
Observation
18
2
7
8
19
15
14
6
20
3
9
5
17
1
10
13
4
12
11
16
41
42
49
43
44
48
38
37
40
36
39
45
46
47
26
22
32
21
27
31
34
28
30
50
35
33
24
25
23
29

Appendix B.3

Number of clusters: 1

	Number of Obs.	Within Sum of Squares	Average 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

Number of clusters: 2

	Number of Obs.	Within Sum of Squares	Average 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	Average 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

Number of clusters: 4

	Number of Obs.	Within Sum of Squares	Average 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	

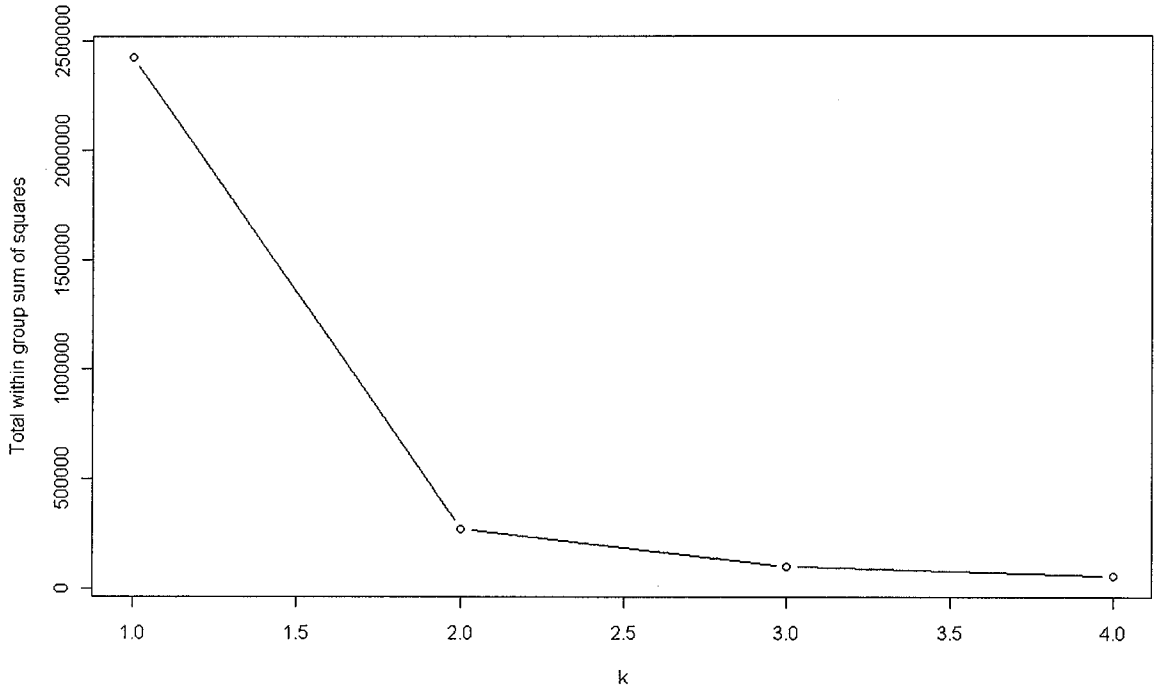
Cluster Centroids:

	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



Appendix C.1

Covariance Matrix:

	age	circumference
age	241930.71	25831.021
circumference	25831.02	3304.891

Principle Components applied on the Covariance Matrix:

Rotation:

	PC1	PC2
age	0.9943232	-0.1064020
circumference	0.1064020	0.9943232

Importance of components:

	PC1	PC2
Standard deviation	494.666	23.2535
Proportion of Variance	0.998	0.0022
Cumulative Proportion	0.998	1.0000

Principle Components applied on the Correlation Matrix:

Rotation:

	PC1	PC2
age	0.7071068	-0.7071068
circumference	0.7071068	0.7071068

Importance of components:

	PC1	PC2
Standard deviation	1.383	0.2941
Proportion of Variance	0.957	0.0432
Cumulative Proportion	0.957	1.0000

Section B - Applied Forecasting

4. **Definitions.**

Write short notes (approx. 150 to 200 words) on FIVE of the following topics (5 marks each):

(a) Regression methods in forecasting.

(b) The definition and fitting of an ARMA(p,q) model.

(c) Autoregression versus indicator variables in seasonal time series models.

(d) The use of the backshift operator in representing ARMA time series models.

(e) The behaviour and implications of the SES model as α approaches its limits.

(f) Transformations to induce stationarity .

(g) The ARMA(p,q) model.

(h) Identifying seasonality.

(25 marks)

5. **Forecasting and model selection**

(a) Explain the Simple Exponential Smoothing (SES) recursion method.

[4 marks]

(b) Explain how the MAPE or RMSE can be used to define the best SES model for a time series.

[4 marks]

(c) Assuming we have the observations $\{y_i\}_{i=1,\cdots,n}$ for which we fitted a linear autoregressive (AR(1)) model such that

$$y_i = \hat{\phi}_0 + \hat{\phi}_1 \, y_{i-1} + \epsilon_i \quad \text{with} \quad \epsilon_i \sim \mathcal{N}(0, s^2), \ \forall i = 2, \cdots, n$$

(i) Forecast y_{n+1} and its 95% prediction interval.

[4 marks]

(ii) Forecast y_{n+k} and its 95% prediction interval.

[4 marks]

(d) The Akaike Information Criterion (AIC) is defined w.r.t. a likelihood L and the number of parameters m involved in a model:

$$AIC = -2\log(L) + 2m \tag{1}$$

(i) Explain how AIC is used in time series analysis. What is the difference with other criteria such as RMSE and MAPE?

[4 marks]

(ii) Equation (1) is sometimes approximated by:

$$AIC \simeq n(1 + \log(2\pi)) + n\log(s^2) + 2m \tag{2}$$

where n is the number of data in the times series. Explain how this approximation (2) can be obtained from equation (1) and which assumptions are used.

[5 marks]

(25 marks)

6. **Times series Analysis.**

(a) Discuss the 3 components that can be found in times series. Use the time series plotted in figure 1 to illustrate your answer.

[4 marks]

(b) Is this time series stationary? Explain your answer.

[4 marks]

(c) Explain how you can infer the period from the ACF of the time series in fig. 1.

[4 marks]

(d) What is a periodogram? How would you use it to compute the period of the time series in figure 1.

[4 marks]

(e) The expert proposes the model expressed using the backshift operator B:

$$(1-B)\,(1-B^{12})\,y_t = (1-\theta_1\,B)\,(1-\Theta_1\,B^{12})\epsilon_t \quad \text{(or } ARIMA(0,1,1)(0,1,1)_{12})$$

(i) Redefine the model $ARIMA(0,1,1)(0,1,1)_{12}$ algebraically (without the backshift operator).

[4 marks]

(ii) Using the plots in figure 2, explain why the expert is proposing this model.

[5 marks]

(25 marks)

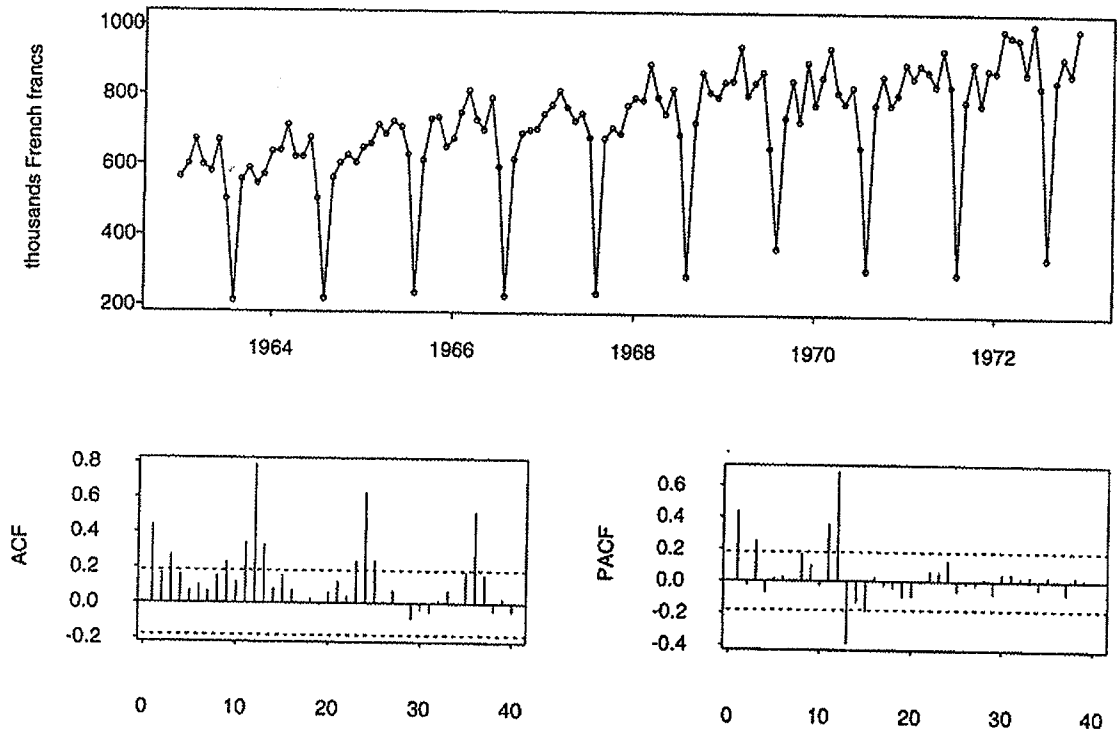


Figure 1: Monthly French Industry sales of printing and writing paper (in thousand francs) between 1963-1972.

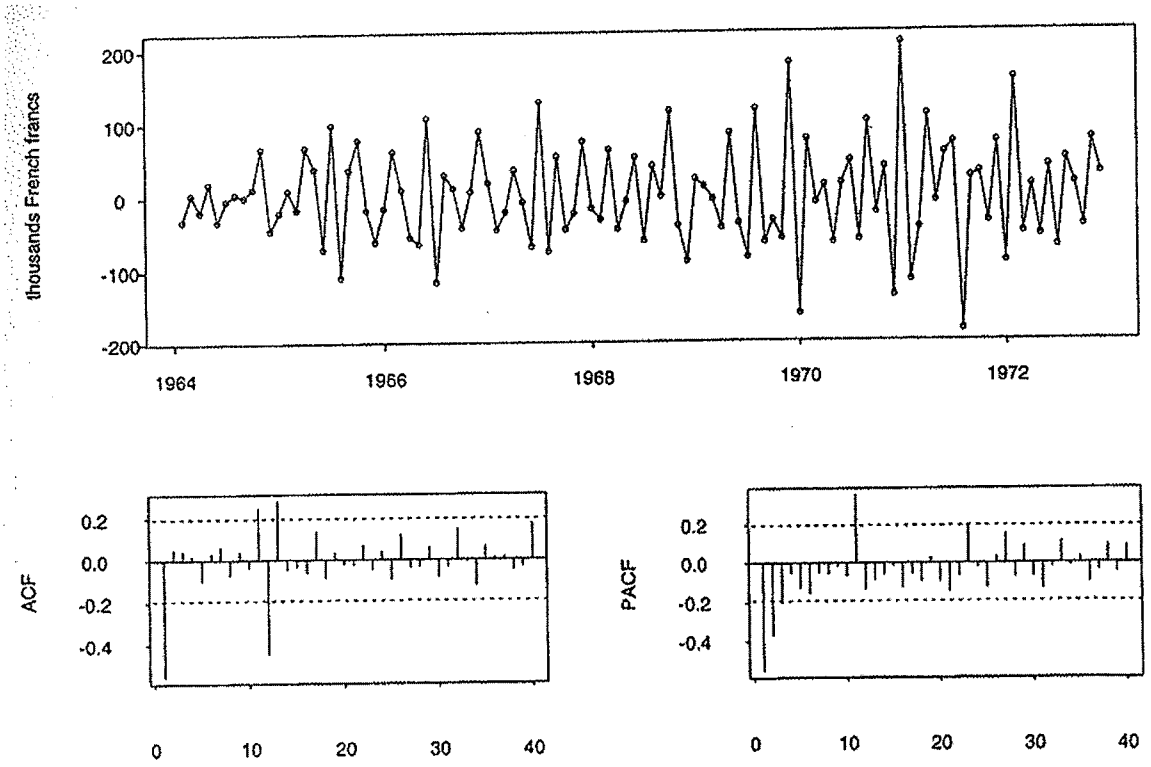


Figure 2: Monthly French Industry sales of printing and writing paper (in thousand francs) between 1963-1972, after seasonal difference and a first difference.