CHAPTER THREE

3.1 STATISTICAL ANALYSIS

3.1.1 Introduction

For investigating the relationship between allergic symptoms and environmental conditions, **a time series analysis** is used, utilizing a STATA package.

In a time series analysis data, that is, sequences of measurements follow non- random orders. The assumption is that successive values in the data file represent consecutive measurements taken at equally spaced time intervals. The data consist of a systematic pattern, and random noise (error). The latter is filtered out to some degree.

Time series patterns describe two components, namely trend and seasonality. Trend represents a general systematic linear or nonlinear component that changes over time and does not repeat, whereas seasonality may have a formally similar nature, but repeats itself in systematic intervals over time.

Two main goals of time series analysis are to identify the nature of the phenomenon represented by the sequence of observations, and forecasting of future values. Once a pattern is established, it can be interpreted and integrated with other data. Extrapolation can then follow to predict future events.

An Autoregressive integrated moving average (ARIMA) analysis procedure is used to provide both moving averages and lag results: a moving average will show vertical fluctuations in the amount of prescriptions for asthma and rhinitis, whilst a lag analysis will show the period between successive fluctuations.

Two common processes are thus in place:

- An autoregressive process (**AR**), where each observation is made up of a random error component, and a linear combination of prior observations.
- A moving average (**MA**) process where each element in the series can also be affected by the past error, that cannot be accounted for by the autoregressive component.

In STATA, ar () and ma () are thus used to specify the lags of autoregressive and moving-average terms respectively.

The Autoregressive Integrated Moving Average model thus includes three parameters in the model, namely the autoregressive parameters, the number of differencing passes, and the moving average parameters.

The **ARIMA** (p, d, q) procedure in STATA is thus used to explain the relationship between a time- dependent outcome variable Y, and a set of k predictor variables X1, X2, …, Xk.

Distributional assumptions must be made about 3 important parameters that determine the pattern of the variation of Y as a function of the k predictor variables through time. These 3 parameters are the number of lags (p), the number of times Y and each of the X's are differentiated (d), and the number of lags of moving averages (q).

Estimation (where the parameters are estimated by means of minimizing the sum of squared residuals) and forecasting (where values are first integrated) are the next steps.

3.1.2 Structural equations for a time series analysis

The structural equation of a time series model containing a first order autoregressive process **AR**(1) and a first order moving average **MA**(1) is given by the following equation:

$\underline{Y_t = X_t \beta + \mu_t}$ where $\mu_t \sim N(0, \delta^2_t)$ (1.1)

In the above expression, the variance δ^2_t is expressed as a function of lagged disturbances. In the above model, it is assumed that the true mean μ_t is normally distributed with mean 0 and constant variance δ^2_t.

If a first order autoregressive AR(1) estimate and a first order moving average MA(1) have to be done, the following structural equation would be suitable:

$Y_t = X_t \beta + \mu_t$ where

$\mu_t = \rho \mu_{t-1} + \theta \varepsilon_{t-1} + \varepsilon_t$ where

ρ = first order autocorrelation parameter

θ = first order moving average parameter

$\varepsilon_t \sim NID(0, \sigma^2)$ are white noise disturbances

In the study and analysis <u>interpretations can thus be made as follows</u>:

Prescriptions for asthma: a time- dependent variable that is affected by the predictor variables rainfall, wind speed, temperature and humidity.

AR (1) (also called the Markov process) indicates that autoregressive terms or lags of order 1 are included in the time series regression model. This means that the process is similar to a multiple linear regression model, but X_t is regressed past values of X_t and not on independent variables. In this study, due to fairly stationary time series

data sets, the order of past values of the predictor variables that affect the dependent variables is only 1.

MA(1 4) indicates that the residuals are assumed to have white noise disturbances, and that a quarterly moving average effect is added. Although the data set in this study is not quarterly data, the time series plots of the set show that it resembles quarterly time series data. In addition to an autoregressive term and a moving average (1) term, a seasonal moving average (4) term at lag 4 is included to account for the remaining quarterly effect.

3.1.3 Evaluation of the model

The following can be looked at to evaluate the model:

1. Parameter estimates for example appropriate t values, are computed from the parameter standard errors.
2. Accuracy of the forecasts: includes a parsimonious model (least parameters and greatest number of degrees of freedom) and the production of statistically independent residuals where the autocorrelogram of the residuals should show no serial dependency between the residuals.
3. Analysis of the residuals: The analysis of the residuals constitutes an important test of the model. They should not be (auto)-correlated, and be normally distributed.
4. Possible bias (As applicable to the study):
   (a) Selection bias:
   - People not taking treatment for disease, presenting late, taking follow up medication for a diagnosis made earlier, or taking over-the-counter prescriptions (not reflected in prescriptions received at GPNet).
   - Late submissions of prescriptions.

   (b) Information Bias:
   - Misclassification of disease (wrong diagnosis) by doctor.
   - Presence of an infective epidemic, e.g. influenza, precipitating further disease.
   - Other climatic variables not measured at present (e.g. pollution).
   - Patients admitted in hospitals, or treated by specialists (not reflected in the database).

3.2 RESULTS OF THE TIME SERIES ANALYSIS

3.2.1 List of variables in the data set

- Year
- Month
- Day
- Rain = quantity of rainfall in mms
- Wind = wind speed in meters per second
- Prh = number of prescriptions for allergic rhinitis (dependant variable)
- Pas = number of prescriptions for asthma (dependant variable)
- Temp = maximum temperature – minimum temperature (degrees Celcius)
- Hum = maximum humidity – minimum humidity (as a %)

The above variables (on a y-axis) are listed against time (365 daily columns, on the x-axis)

3.2.2 Calculation

403 observations were used, as well as the following variables namely year, month, day, rain, wind, prescriptions for asthma, prescriptions for allergic rhinitis, temperature and humidity.

Time was then generated, a t –set done, and an **ARIMA** process done namely

ARIMA pas rain wind temp humidity ar (1) ma (1 4).

Parameter estimates were obtained by Maximum Likelihood estimation; an **ARIMA** regression was then done (See Table one for values).

Part 1:

ARIMA regression for prescriptions for asthma

Sample: 01mar2002 to 07apr2003 n = 403

Wald chi2(7) = 453.95

Log likelihood = -800.0464 Prob > chi^2 = 0.0000

Since STATA does not give multiple R-squared values for ARIMA procedures, SPSS calculations were done to determine the Akaike Information Criterion (AIC) and the Schwarz Bayesian Criterion (SBC). The small values of 1228.3898 and 1256.365 obtained respectively, indicated that the fitted models are parsimonious.

Table 1 Parameter estimates obtained from a Time Series Regression

pas	Coefficient.	Standard Error.	z	P>|z|	95% Confidence Interval	
rain	.000446	.0010305	0.43	0.665	-.0015737	.0024656
wind	-.000366	.0008255	-0.44	0.658	-.0019839	.001252
temp	-.0011625	.0005927	-1.96	0.050	-.0023241	-8.95
hum	.0004675	.0002817	1.66	0.097	-.0000847	.0010197
cons	1.195743	.1128314	10.60	0.000	.9745977	1.416889
ARIMA						
L1 (Lag#1)	.7674467	.0812798	9.44	0.0000	.6081413	.9267522
L1	-.8103576	.0826385	-9.81	0.000	-.9723262	-.6483891
L4 (Lag#4)	-.0895653	.0475581	-1.88	0.060	-.1827775	.0036469
/sigma	1.761126	.1017455	17.31	0.000	1.561708	1.960543

Interpretation of results from the regression of prescriptions for Asthma on rainfall, wind speed, temperature range and humidity.

The regression coefficients at lags 1 and 4 are different from each other. The results show that the regression coefficients of AR and MA are significant at lag number 1, but not at lag number 4. This is to be expected, due to the fairly stationary time series data.

Prescriptions for asthma was regressed on rain, wind, temp and humidity.

An α- value of 10% represents the maximum probability of making a Type-I error, viz. wrongfully rejecting a parameter under the null hypothesis. In the context of an

AR- model, as used in this dissertation, α denotes the first-order parameter in the model. In practice the estimate of this parameter is usually 0.30 or lower.

A p-value, or probability of obtaining a result as extreme (or more) than the one observed (if the null hypothesis is true), was obtained after the statistical test has been performed. Estimated parameters with p-value <0.05 are deemed statistically significantly different from zero, and should remain in the model.

At the α = 10% level of significance:

Rainfall does not influence prescriptions for asthma since $p = 0.665 > \alpha$.

Wind speed does not influence prescriptions for asthma since $p = 0.658 > \alpha$.

Temperature influences prescriptions for asthma since $p = 0.050 < \alpha$.

Humidity influences prescriptions for asthma since $p = 0.097 < \alpha$.

Concerning **residuals:**

Residuals are the differences between estimated results and true values. Large values will indicate incorrect estimations, and small values of the residuals indicate accurate estimations. STATA standardizes the residuals, and plots a normal probability plot of residuals. Ideally it should be a straight line, but if it resembles an S- shape (as in this study: see probability plot), it means that the white noise assumption is satisfied, and hence the time series has achieved a state of equilibrium.

$\hat{\beta}_{AR(L1)} = 0.7674467$ with $p = 0.000 < \alpha = 0.05$.

Since $p < \alpha$, there is a statistically significant first order autocorrelation in the disturbances.

$\hat{\beta}_{MA(L1)} = -0.8103576$ with $p = 0.000 < \alpha = 0.05$.

Since $p < \alpha$, there is a statistically significant first order moving average.

$\hat{\beta}_{MA(L4)} = -0.0895653$ with $p = 0.060 > \alpha = 0.05$.

Since $p > \alpha$, there is no significant fourth order moving average.

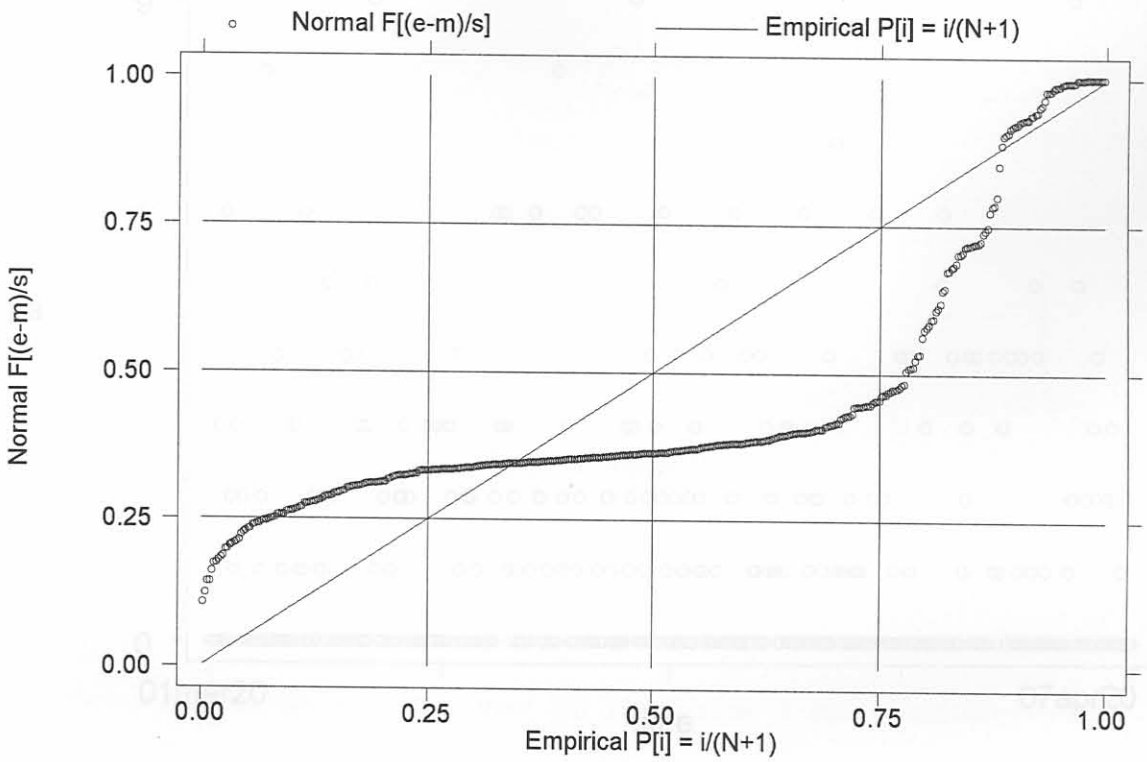

Figure 1 Normal Probability Plot of Residuals

The S- shaped normal probability plot indicates a violation of the normal assumption of the error structure. The S- shape indicates that a distribution with lighter tails than the normal distribution fits the residual series.

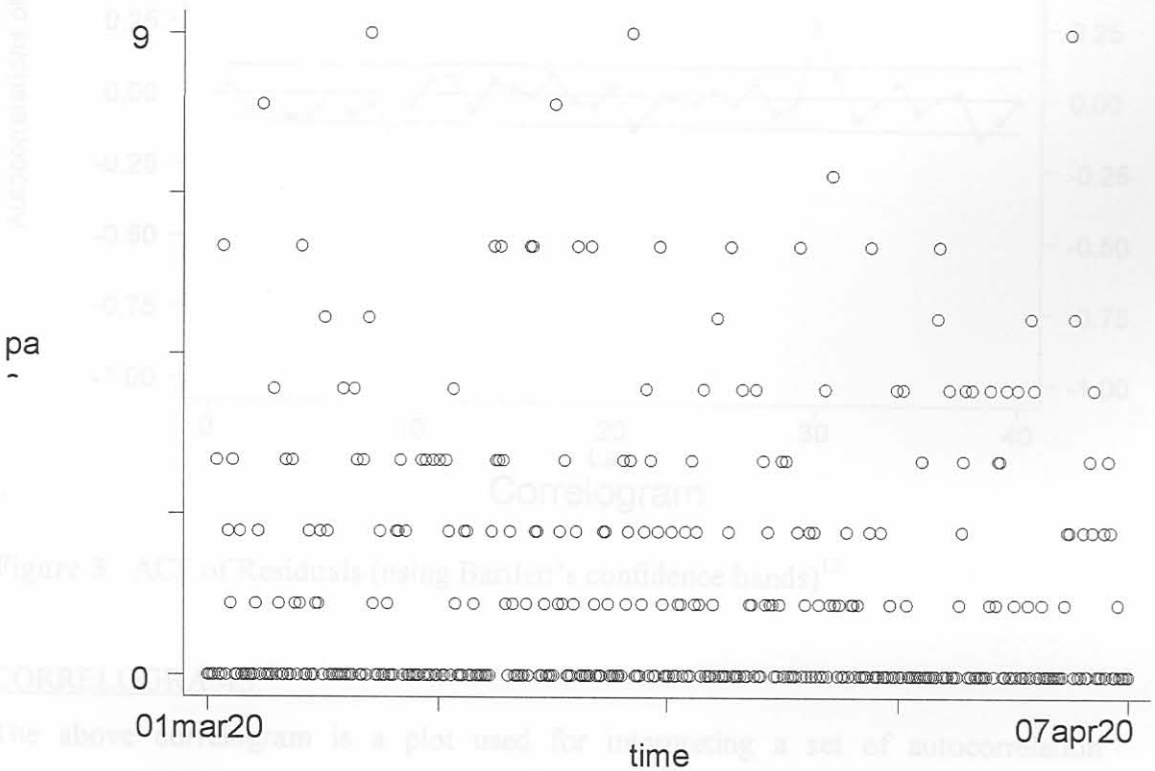

Figure 2 Time Series Plot of Asthma Prescriptions

The above graph shows the trend followed by prescriptions for asthma as time varied from 01 March 2002 to 07 April 2003. (Y- axis shows amount of prescriptions on a daily basis, from 0 to 9; X- axis shows the time, divided into 4 quarters).

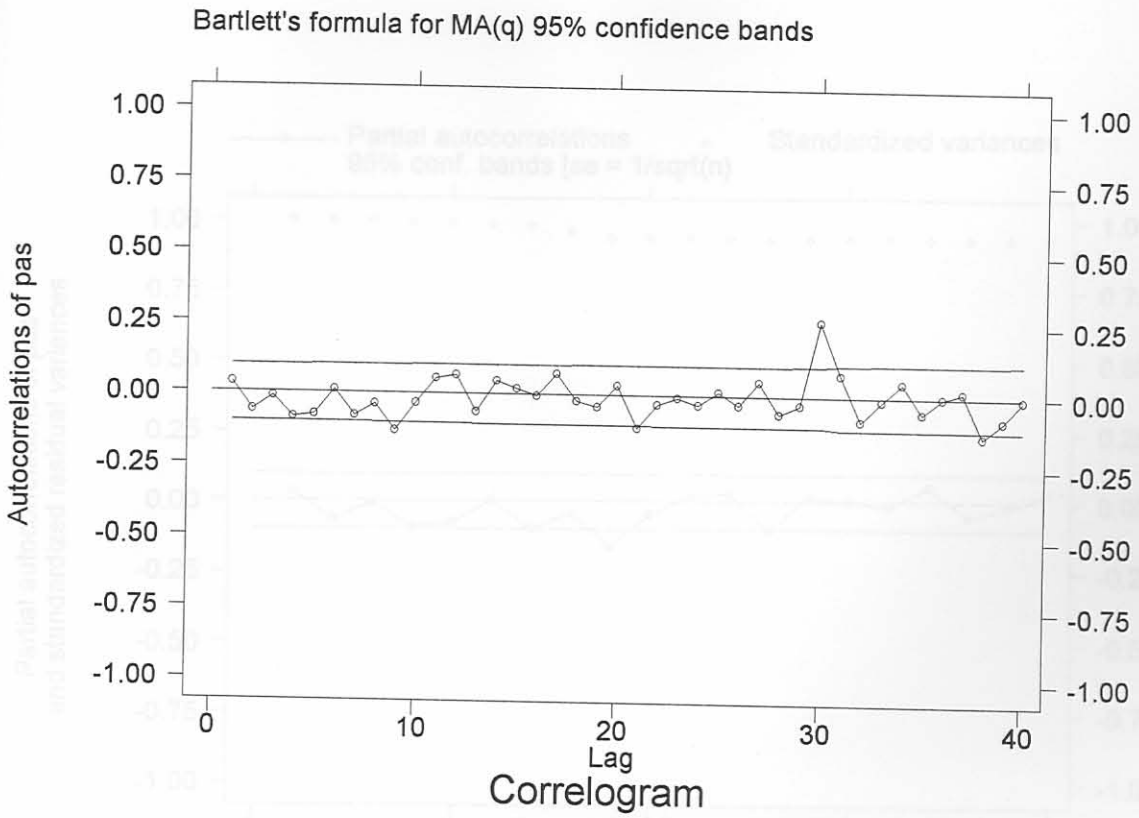

Figure 3 ACF of Residuals (using Bartlett's confidence bands)[13]

CORRELOGRAMS:

The above correlogram is a plot used for interpreting a set of autocorrelation coefficients. It is a graph of r_k versus lag k.

For a large sample, r_k is approximately normally distributed with expected value 0 and variance 1/N. An approximate 95% confidence interval for r_k is therefore given by [-2/sqrt(N); 2/sq(N)]. In this study, N=403, and an approximate 95% confidence interval for r_k is hence given by [-0.0996; 0.0996].

40 Lag periods were used: 0 staring at the start of study period, and 40 denoting the end of study period.

The above correlogram shows a fairly stationary time series pattern for prescriptions for asthma, with a break-out at lag number 30, reflecting an increase in prescriptions during the rainy season (November and December).

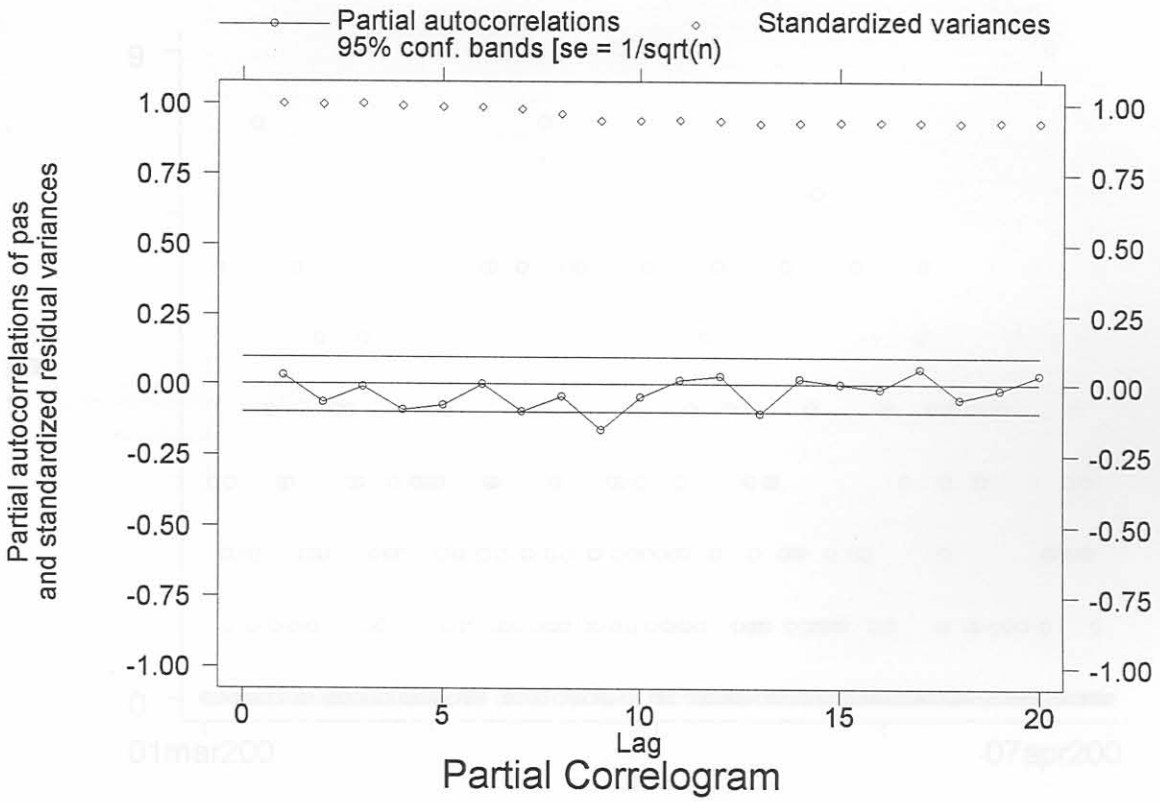

Figure 4 PACF of Residuals

PARTIAL CORRELOGRAMS:

The above partial Correlogram gives a plot of partial autocorrelations and standardized residual variances versus lag periods, very similar to the correlogram. It is however more precise, as it gives a 95% confidence band for autocorrelations. Again it is shown that the time series for prescriptions for asthma is fairly constant, with no outliers.

20 lag periods were used in the above to rule out the presence of unexpected significant results at higher lag periods.

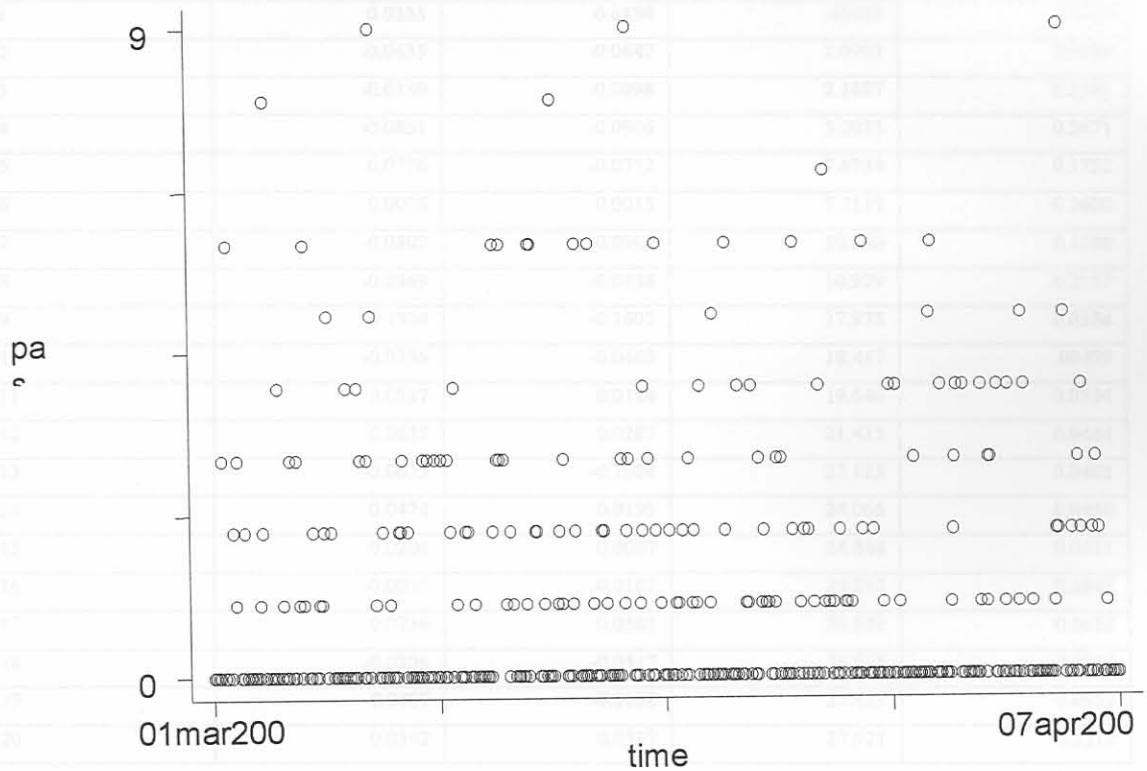

Figure 5 Time Series Plot of Asthma Prescriptions

Table 2 : Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) for 20 lags

LAG	Auto Correlation	Partial Auto Correlation	Prob>Q [Autocorrelation]	[Partial Autocor]
1	0.0333	0.0334	.45083	0.5019
2	-0.0635	-0.0647	2.0903	0.3516
3	-0.0139	-0.0098	2.1687	0.5381
4	-0.0861	-0.0906	5.2033	0.2671
5	-0.0776	-0.0752	7.6734	0.1752
6	0.0096	0.0015	7.7117	0.2600
7	-0.0802	-0.0967	10.366	0.1688
8	-0.0369	-0.0438	10.929	0.2057
9	-0.1304	-0.1605	17.978	**0.0354**
10	-0.0336	-0.0460	18.447	**00479**
11	0.0537	0.0144	19.646	0.0504
12	0.0655	0.0287	21.435	**0.0444**
13	-0.0635	-0.1024	23.123	**0.0402**
14	0.0474	0.0199	24.066	**0.0450**
15	0.0206	0.0007	24.244	0.0611
16	-0.0030	-0.0167	24.247	0.0842
17	0.0736	0.0563	26.539	0.0652
18	-0.0206	-0.0517	26.718	0.0844
19	-0.0407	-0.0198	27.423	0.0952
20	0.0342	0.0337	27.921	0.1113

Discussion:

The above table shows values of autocorrelations and partial autocorrelations for each of the 20 lags assumed for data analysis. The Q statistic is used to measure the strength of autocorrelations (AC) and partial autocorrelations (PAC) at each lag. If the p-value next to the Q statistic falls below 0.05, it means there is a significant AC and PAC at the lag. (A p-value greater than 0.05 means no significant AC and PAC at that specific lag).

The p-values for lags 9, 10, 12, 13 and 14 (printed in bold, and reflecting the time period August 2002 to December 2002) are each less than 0.05, thus significant at the 5% level of significance. It also correlates with the rainy season in Pretoria. This shows that the autocorrelations and partial correlations are statistically significant at lags 9, 10, 12, 13 and 14 at the 5% level of significance.

PREDICTING FUTURE VALUES OF PRESCRIPTIONS FOR ASTHMA (PAS)

The estimated ARIMA model for prescriptions for asthma as a function of rain, wind, temperature and humidity is given by the following equation:

Pas= 1.195743 + 0.000446 x rain - 0.000366 x wind - 0.0011625 x temperature + 0.0004675x humidity (1.2)

SCATTER PLOT FOR PRESCRIPTIONS FOR ASTHMA (PAS) OVER TIME

n = 403

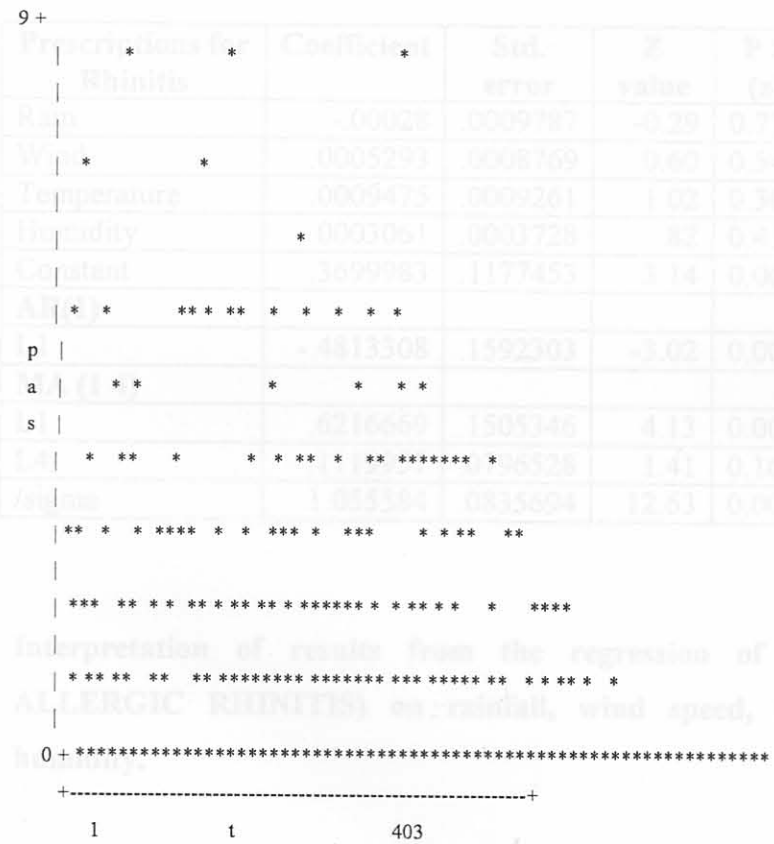

Figure 6 Time Series Plot of Predicted Asthma Prescriptions (PAS)

Part 2:

ARIMA regression for prescriptions for allergic rhinitis

Sample: 01mar2002 to 07apr2003 n = 178

Wald chi^2 (7) = 82.53

Log likelihood = -593.689 Prob > chi^2 = 0.0000

| | Semi-robust

prh | Coef. Std. Err. z P>|z| [95% Conf. Interval]

Table 3 Parameter estimates obtained from a Time Series Regression

Prescriptions for Rhinitis	Coefficient	Std. error	Z value	P > (z)	95% Conf.Interval	
Rain	-.00028	.0009787	-0.29	0.775	-.0021982	.0016382
Wind	.0005293	.0008769	0.60	0.546	-.0011894	.0022481
Temperature	.0009475	.0009261	1.02	0.306	-.0008676	.00027626
Humidity	.0003061	.0003728	.82	0.412	-.0004246	.0010367
Constant	.3699983	.1177453	3.14	0.002	.1392218	.6007749
AR(1)						
L1	-.4813308	.1592303	-3.02	0.003	-.7934164	-.1692452
MA (1 4)						
L1	.6216669	.1505346	4.13	0.000	.3266246	.9167093
L4	.1119957	.0796528	1.41	0.160	-.044121	.2681124
/sigma	1.055584	.0835694	12.63	0.000	.8917905	1.219377

Interpretation of results from the regression of prh (prescriptions for ALLERGIC RHINITIS) on rainfall, wind speed, temperature range and humidity.

At the α = 10% level of significance,

Rainfall doesn't influence prescriptions for allergic rhinitis, since p=0.775 > α.

Wind speed does not influence prescriptions for allergic rhinitis, since p = 0.546 > α.

Temperature does not influence prescriptions for allergic rhinitis since p = 0.306 > α.

Humidity does not influence prescriptions for allergic rhinitis, since $p = 0.412 > \alpha$.

Examining the residuals:

$\hat{\beta}_{AR(L1)} = -0.4813308$ with $p = 0.003 < \alpha = 0.05$.

Since $p < \alpha$, there is a statistically significant first order autocorrelation in the disturbances.

$\hat{\beta}_{MA(L1)} = 0.6216669$ with $p = 0.000 < \alpha = 0.05$.

Since $p < \alpha$, there is a statistically significant first order moving average.

$\hat{\beta}_{MA(L4)} = 0.1119957$ with $p = 0.160 > \alpha = 0.05$.

Since $p > \alpha$, there is no significant fourth order moving average.

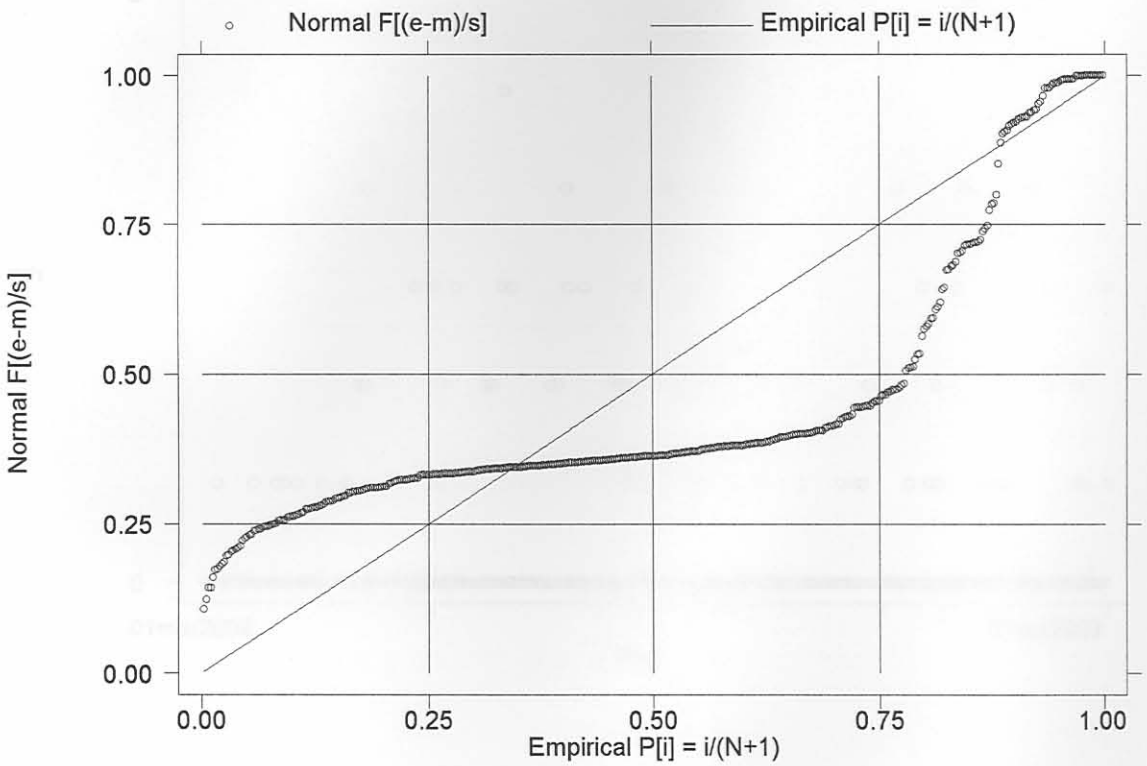

Figure 7 Normal Probability Plot of Residuals

The S- shaped normal probability plot indicates a violation of the normal assumption of the error structure. The S- shape indicates that a distribution with lighter tails than the normal distribution fits the residual series.

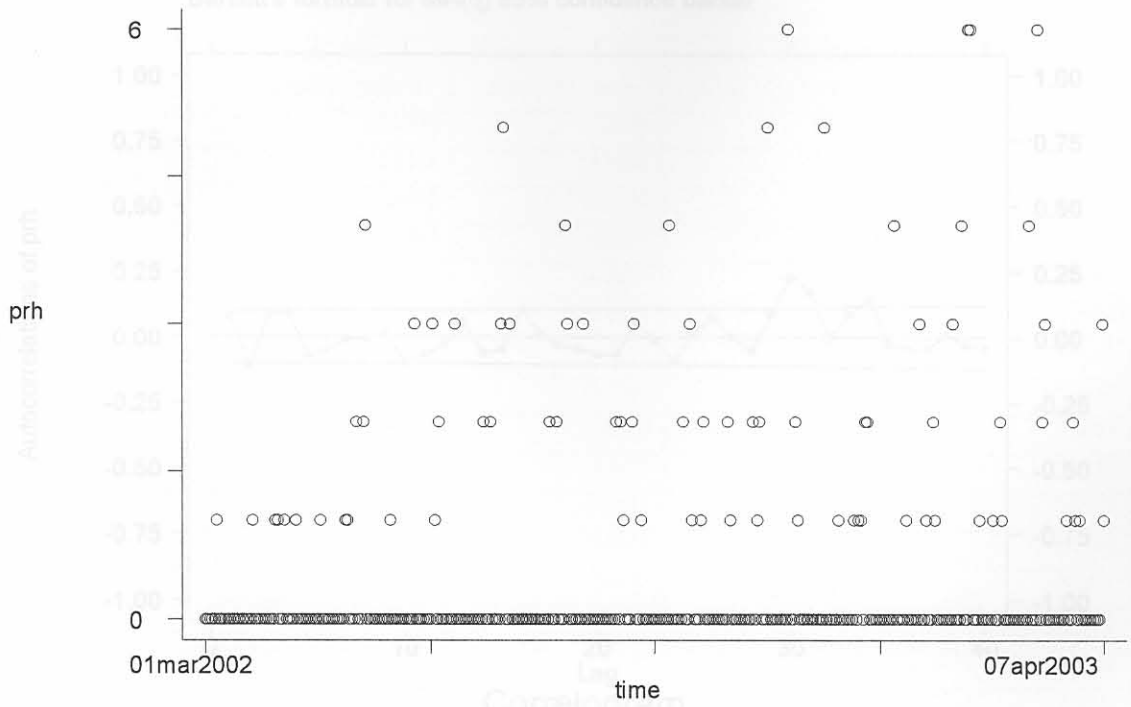

Figure 8 Time Series Plot of Allergic Rhinitis Prescriptions

The above graph shows the trend followed by prescriptions for allergic rhinitis as time varied from 01 March 2002 to 07 April 2003. (Maximum of 6 prescriptions / day noted)

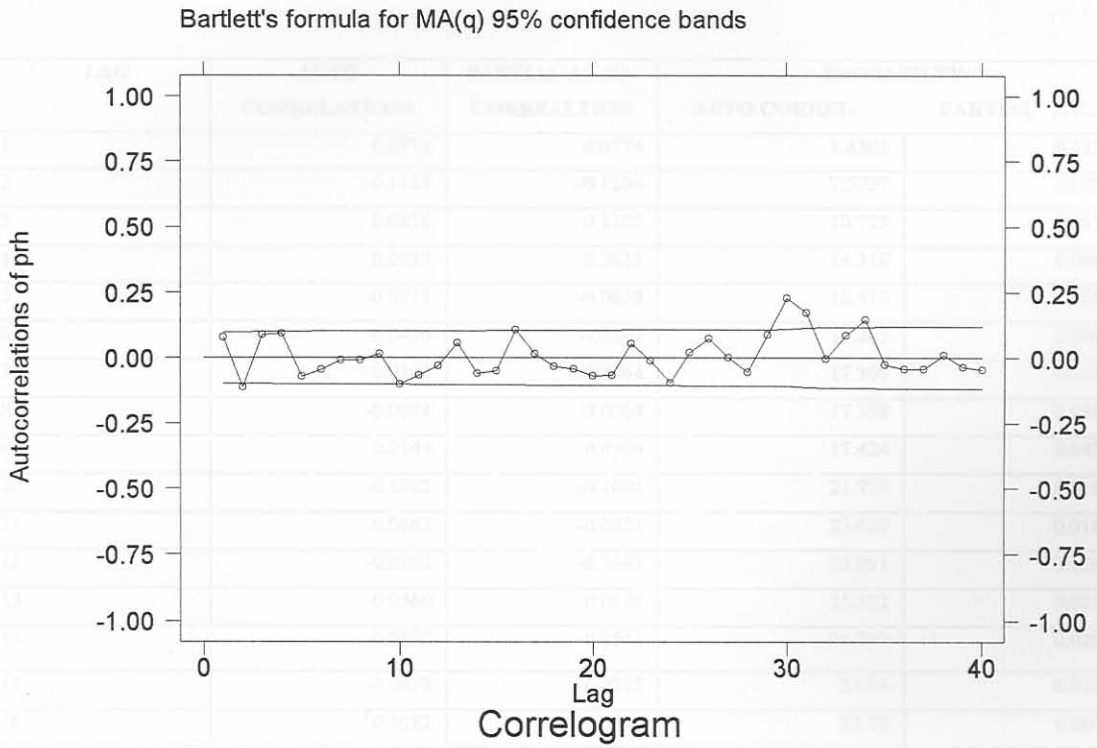

Figure 9 ACF of Residuals (using Bartlett's confidence bands)

The above graph is a plot of autocorrelations, again showing breakout during the rainy season

Table 4: ACF and PACF for 20 lags

LAG	AUTO CORRELATIONS	PARTIAL AUTO-CORREALTION	PROBABILTY	
			AUTO CORREL.	PARTIAL A.C.
1	0.0774	0.0774	2.4305	0.1190
2	-0.1125	-0.1206	7.5797	0.0226
3	0.0878	0.1105	10.725	0.0133
4	0.0937	0.0653	14.318	0.0063
5	-0.0715	-0.0658	16.412	0.0058
6	-0.0456	-0.0249	17.267	0.0084
7	-0.0101	-0.0364	17.309	0.0155
8	-0.0084	-0.0064	17.338	0.0268
9	0.0144	0.0304	17.424	0.0425
10	-0.1022	-0.1095	21.759	0.0164
11	-0.0667	-0.0451	23.609	0.0145
12	-0.0302	-0.0543	23.991	0.0204
13	0.0560	0.0679	25.302	0.0211
14	-0.0597	-0.0541	26.797	0.0205
15	-0.0498	-0.0275	27.84	0.0226
16	0.1082	0.0943	32.78	0.0079
17	0.0149	-0.0261	32.874	0.0117
18	-0.0332	0.0064	33.341	0.0152
19	-0.0442	-0.0608	34.17	0.0175
20	-0.0699	-0.1076	36.254	0.0144

The above table shows values of autocorrelations and partial autocorrelations for each of the 20 lags assumed for data analysis. The p-values for lags 2 to 20 are each less than 0.05. This shows that the autocorrelations and partial correlations are statistically significant at lags 2 to 20 at the 5% level of significance.

Figure 10 Time Series Plot of Predicted Allergic Rhinitis Prescriptions

The above graph shows a scatter plot of prescriptions for allergic rhinitis versus time, prescriptions varying between 0 to maximum 6 per day.

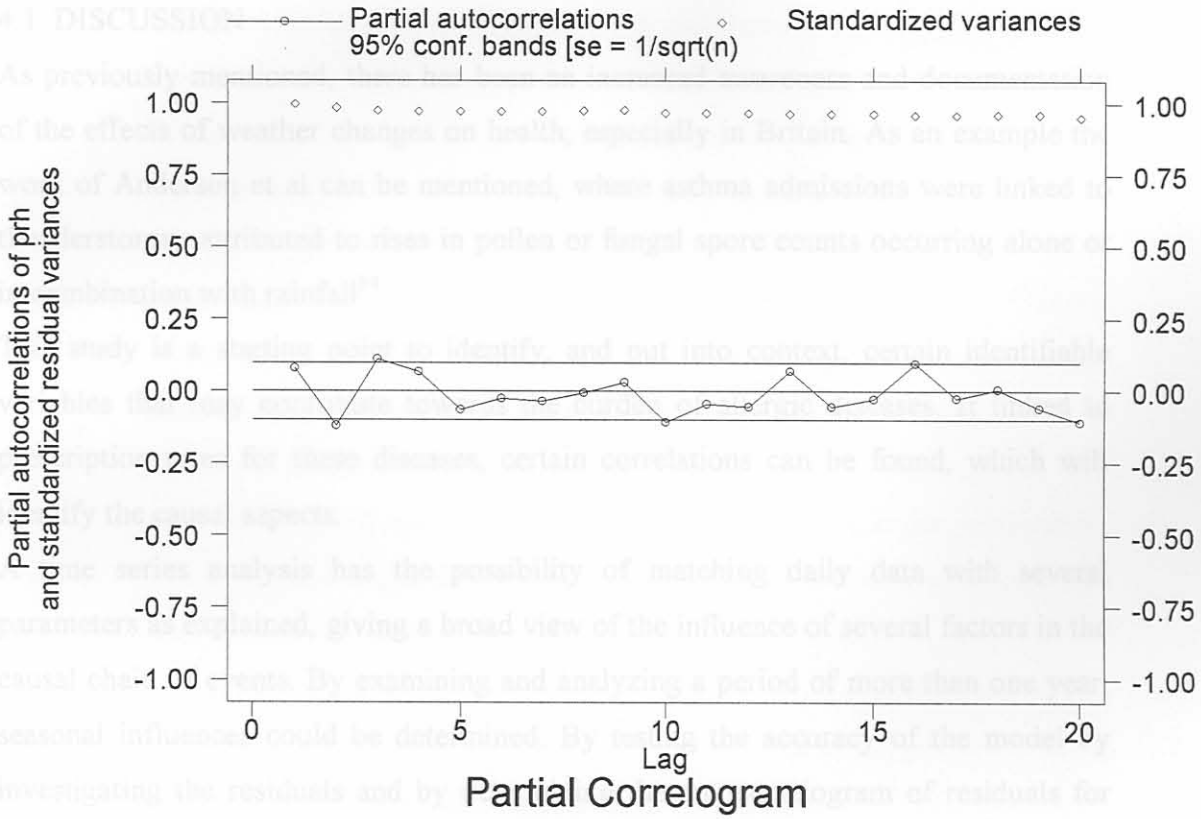

Figure 11 PACF of Residuals

The above graph shows a plot of partial autocorrelations versus lags for prescriptions for allergic rhinitis (prh). It can be seen from the correlogram plot that values of prescriptions for allergic rhinitis fluctuated during the period of study.