

RESPIRATORY DISEASES AND VIRAL INFECTIONS

TESTING THE RELATIONSHIP BETWEEN RESPIRATORY DISEASES
AND VIRAL INFECTIONS IN VARIOUS AGE GROUPS

By
LEANNE SANTARELLI, B.Sc.

A Thesis
Submitted to the School of Graduate Studies
in Partial Fulfillment of the Requirements
for the Degree
Master of Science

McMaster University
©Copyright by Leanne Santrelli, December, 2004

MASTER OF SCIENCE (2004)
(Statistics)

McMaster University
Hamilton, Ontario

TITLE: Testing the Relationship Between Respiratory Diseases
 and Viral Infections in Various Age Groups
AUTHOR: Leanne Santarelli, B.Sc. (McMaster University)
SUPERVISOR: Dr. Aaron Childs
NUMBER OF PAGES: viii, 64

Abstract

The objective of this project was to investigate and determine the association between hospitalizations of respiratory diseases with one another and with isolations of viral infections in five age groups. Weekly data on all hospitalizations in Ontario, Canada, from week 14 of 2001 to week 13 of 2003 were obtained for five age groups (under 2 years, 2 to 4 years, 5 to 15 years, 16 to 49 years and over 50 years inclusive) for respiratory diseases including, asthma, respiratory tract infection (RTI) and chronic obstructive pulmonary disease (COPD)¹. Furthermore, data for viral infections including influenza virus type A and type B (Flu AB) and respiratory syncytial virus (RSV) isolations were also obtained from Health Canada for the same weekly time periods.

In order to test for independence and determine a relationship, if any, between hospitalizations of respiratory diseases with one another and with isolations of viral infections, structural time series models were developed for all age groups of the respiratory diseases and explanatory variables were modeled accordingly against the hospital admission counts for the respiratory diseases. These explanatory variables include, other respiratory diseases, viral infections, and lagged values of the dependent variable. Neither FLU AB nor RSV showed a significant relationship with asthma patients of all ages. Weekly RSV peaks coincided with RTI patients under 2 years and RTI peaks in patients 5 to 15 years preceded FLU AB peaks. A relationship between all three respiratory diseases, asthma RTI and COPD, was discovered for all age groups. Peaks of asthma coincided with various transformations of RTI peaks for the five age groups and peaks of COPD coincided with both the untransformed asthma and RTI peaks in patients over 50. For all other relationships, the null hypothesis of independence was accepted. These findings suggest that there is a strong association between respiratory diseases and that children and adults with respiratory diseases respond differently to viral infections.

¹Only data for patients over 50 years was obtained for COPD.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Dr. Aaron Childs, for his constant and excellent guidance throughout my project. I would also like to sincerely thank Mr. Neil Johnston from the Firestone Institute for Respiratory Health at St. Joseph's Hospital for giving me the opportunity to work on this project and providing me with the data. I would also like to extend my thanks to Neil and Jennifer Dai for meeting with me and Dr. Childs throughout the course of my analysis and providing insightful discussions and ideas. Finally, I would like to extend my gratitude to my family and friends for their constant support and encouragement.

Table of Contents

| | |
|---|-------------|
| Abstract | ii |
| Acknowledgements | iii |
| List of Tables | vii |
| List of Figures | viii |
| 1 Introduction | 1 |
| 1.1 Motivation | 1 |
| 1.2 Introduction to Time Series | 3 |
| 1.3 Cross-Correlation Approach for Testing the Independence of Two Series | 5 |
| 2 Structural Time Series | 7 |
| 2.1 State-Space Representation | 7 |
| 2.2 The Kalman Filter | 8 |
| 2.3 Estimation for State-Space Models | 9 |
| 2.4 Predictor Variables and Lagged Dependent Variables | 11 |
| 3 Methodology | 13 |
| 3.1 General Structural Time Series Model | 13 |
| 3.2 Diagnostic Tests | 15 |
| 3.3 Goodness of Fit | 18 |
| 3.4 Model Selection | 19 |
| 4 Modeling Example | 21 |
| 5 Results and Discussion | 32 |
| 5.1 Asthma Age Groups | 33 |
| 5.2 RTI Age Groups | 34 |
| 5.3 COPD Over 50 | 35 |
| 5.4 Respiratory Disease Analysis By Age Group | 35 |
| 5.5 Discussion | 36 |
| 6 Conclusions and Future Work | 38 |

| | |
|------------------------------------|-----------|
| Appendix | 40 |
| Appendix A: SAS Program | 40 |
| Appendix B: STAMP Output | 58 |
| Bibliography | 64 |

List of Tables

| | | |
|------|---|----|
| 1.1 | Summary of Asthma Age Groups and Respiratory Viruses | 2 |
| 1.2 | Summary of RTI Age Groups and Respiratory Viruses | 2 |
| 1.3 | Summary of Asthma and RTI Age Groups Viruses | 2 |
| 1.4 | Summary of COPD | 2 |
| 4.1 | Fit Statistics of ASU2 Model | 25 |
| 4.2 | Fit Statistics of log(ASU2) Model | 25 |
| 4.3 | Parameter Estimates of Model 1 | 25 |
| 4.4 | Significance Analysis of Components of Model 1 | 26 |
| 4.5 | Fit Statistics of Model 2 | 26 |
| 4.6 | Parameter Estimates Model 2 | 27 |
| 4.7 | Significance Analysis of Components of Model 2 | 27 |
| 4.8 | Fit Statistics of Model 3 | 27 |
| 4.9 | Parameter Estimates Model 3 | 27 |
| 4.10 | Significance Analysis of Components of Model 2 | 27 |
| 4.11 | Ljung-Box Residual Test | 29 |
| 4.12 | Other Residual Tests | 29 |
| 4.13 | RSV Predictors | 30 |
| 4.14 | FLU AB Predictors | 30 |
| 5.1 | Asthma Age Group Models with Viral Infections (RSV FLU AB) as Predictors . . . | 33 |
| 5.2 | RTI Age Group Models with Viral Infections (RSV FLU AB) as Predictors | 34 |
| 5.3 | COPD Over 50 Model with Viral Infections (RSV FLU AB) as Predictors | 35 |
| 5.4 | Respiratory Disease Analysis By Age Group | 35 |

List of Figures

| | | |
|-----|--|----|
| 4.1 | <i>Plot of asthma patients under 2 years and RSV positive tests (over mean)</i> | 21 |
| 4.2 | <i>Plot of asthma patients under 2 years and FLU AB positive tests (over mean)</i> | 22 |
| 4.3 | <i>Cross Correlation Plot of transformed ASU2 and RSV data</i> | 23 |
| 4.4 | <i>Cross Correlation Plot of transformed RSV and ASU2 data</i> | 23 |
| 4.5 | <i>Cross Correlation Plot of transformed ASU2 and FLU AB data</i> | 24 |
| 4.6 | <i>Cross Correlation Plot of transformed FLU AB and ASU2 data</i> | 24 |
| 4.7 | <i>Residual Plot of ASU2 model (Model 3)</i> | 28 |
| 4.8 | <i>ACF of the residuals of the ASU2 model (Model 3)</i> | 29 |
| 5.1 | <i>COPD and Asthma Patients Over 50 (over mean)</i> | 36 |
| 5.2 | <i>COPD and RTI Patients Over 50 (over mean)</i> | 36 |

Chapter 1

Introduction

1.1 Motivation

Predictable cycles of respiratory diseases requiring hospital treatment occur globally and coincide with peaks of isolations of certain viral infections. For example, influenza is believed to increase the likelihood of asthma and COPD exacerbations, but the specific affects on asthma in children and asthma and COPD in adults is unknown. Furthermore, influenza vaccination may offer protection to asthmatics, but this by no means is a certainty and the affects of vaccination on young children are unknown. Thus, knowing the relationship between disease and virus peaks can provide insight as to how these respiratory diseases can be controlled and hospitalizations prevented.

Weekly data on the number of hospitalizations in Ontario due to respiratory diseases (asthma, RTI and COPD) were collected from week 14 of 2001 to week 13 of 2003 inclusive. The number of patients under 2 years, 2 to 4 years, 5 to 15 years, 16 to 49 years and over 50 years admitted to hospitals for clinically diagnosed asthma or RTI were recorded weekly, along with the number of patients over the age of 50 diagnosed with COPD. These age groups were chosen in order to investigate the different affects viral infections have on patients of different ages and compare them with what is commonly believed. Children under 5 years of age do not have fully developed immune systems, are not in school and children under 2 years are very susceptible to RSV. Children 5-15 years are “school-aged” children and are shown to be the principle source of RTI. Adults 16-49 years show different patterns of respiratory diseases than those patients over 50 years. Furthermore, the two respiratory viruses of concern, FLU AB and RSV were recorded as the number of positive virus

tests at a specific time t .

The purpose of this study is to test the independence of respiratory diseases and viral infections in the 5 age groups as well as the independence of the different respiratory diseases themselves in the same age groups. Similar to the approach and methods used by Scuffham in 2003 and 2004, structural time series is used with the statistical programs SAS and Structural Time series Analyses, Modeler and Predictor (STAMP) to determine if any significant relationships exist. The results of the analyses are found below in Table 1.1, Table 1.2, Table 1.3 and Table 1.4.

Table 1.1: Summary of Asthma Age Groups and Respiratory Viruses

| Asthma Ages (years) | RSV | Flu AB |
|---------------------|------|--------|
| Under 2 | N.S. | N.S. |
| 2-4 | N.S. | N.S. |
| 5-15 | N.S. | N.S. |
| 16-49 | N.S. | N.S. |
| Over 50 | N.S. | N.S. |

Table 1.2: Summary of RTI Age Groups and Respiratory Viruses

| RTI Ages (years) | RSV | Flu AB |
|------------------|--------|--------|
| Under 2 | 0.0032 | N.S. |
| 2-4 | N.S. | N.S. |
| 5-15 | N.S. | 0.0016 |
| 16-49 | N.S. | N.S. |
| Over 50 | N.S. | N.S. |

Table 1.3: Summary of Asthma and RTI Age Groups Viruses

| Asthma and RTI Ages (years) | p-value |
|-----------------------------|--------------|
| Under 2 | 2.04916 E-9 |
| 2-4 | 0.0008 |
| 5-15 | 0.0104 |
| 16-49 | 0.0128 |
| Over 50 | 8.74034 E-86 |

Table 1.4: Summary of COPD

| Analysis | p-value |
|-------------------------|---------|
| COPD and Asthma Over 50 | 0.0000 |
| COPD and RTI Over 50 | 0.0000 |
| COPD and RSV | N.S. |
| COPD and FLU | N.S. |

N.S. : not significant

1.2 Introduction to Time Series

In this section the basic ideas of time series analysis are introduced based on the book, *Introduction to Time Series and Forecasting*, by Brockwell and Davis, 2002. A particular look at concepts of stationarity, by transforming the data to remove trend and seasonal components, and the autocovariance and autocorrelation functions will be discussed.

A *time series*, say $\{X_t\}$, is a set of observations X_t , each being recorded at a specific time t . A *discrete time series*, the type focused on throughout this report, is a time series in which the set T_0 of times at which observations are made is a discrete set.

The **mean function** of $\{X_t\}$ is defined as

$$\mu_X(t) = E(X_t), \text{ where } E[X_t^2] < \infty$$

and the **covariance function** of X_t is defined as

$$\gamma_X(r, s) = \text{Cov}(X_r, X_s) = E[(X_r - \mu_X(r))(X_s - \mu_X(s))]$$

for integers r and s . A time series is said to be stationary if

1. $\mu_X(t)$ is independent of t ,
2. $\gamma_X(t+h, t)$ is independent of t for each h .

Finally, for a stationary time series X_t , the **autocovariance function (ACVF)** of X_t is,

$$\gamma_X(h) = \gamma_X(h, 0) = \gamma_X(t+h, t) = \text{Cov}(X_{t+h}, X_t),$$

and the **autocorrelation function (ACF)** of $\{X_t\}$ is,

$$\rho_X(h) \equiv \frac{\gamma_X(h)}{\gamma_X(0)} = \text{Cor}(X_{t+h}, X_t)$$

Often time series are non-stationary due to strong dependence of variability on the level of the series along with a trend and seasonal components in the data, thus these components should be reduced or eliminated to make the series stationary. In order to remove the dependence of variability and trend and seasonal components, the Box-Cox transformation and differencing techniques should be applied. A general variance-stabilizing transformation is the Box-Cox transformation f_λ and is defined as:

$$f_\lambda(U_t) = \begin{cases} \lambda^{-1}(U_t^\lambda - 1), & U_t \geq 0, \lambda > 0, \\ \ln U_t, & U_t > 0, \lambda = 0. \end{cases}$$

Trend and seasonal components can be detected by examining the graph of the series and also can be identified by autocorrelation functions that are slowly decaying and/or nearly periodic. Trend

and seasonality can be eliminated by differencing. In differencing, the backward shift operator B is defined by $BX_t = X_{t-1}$. The lag- d difference operator ∇^d is often used to eliminate trend and is defined by,

$$\nabla^d X_t = (1 - B)^d X_t.$$

On the other hand, the lag- d difference operator, ∇_d , is used to eliminate seasonal components with period d and is defined by,

$$\nabla_d X_t = (1 - B^d) X_t.$$

The transformations mentioned above allow for stationarity of the series which is necessary for fitting an appropriate ARMA model to the data with zero mean, when the method of structural time series modeling (discussed in Chapter 2) is not being used. The most common models used to fit a stationary series are: an autoregressive process of order p , $AR(p)$ model, a moving average process of order q , $MA(q)$ model, and a mixture of both an $AR(p)$ process and the $MA(q)$ process, referred to as an $ARMA(p, q)$ model. The above models are defined as follows;

1. $\{X_t\}$ is an **autoregressive process of order p** , ($AR(p)$) if

$$X_t = \phi_1 X_{t-1} + \dots + \phi_p X_{t-p} + Z_t$$

2. $\{X_t\}$ is an **moving average process of order q** , ($MA(q)$) if

$$X_t = Z_t + \theta_1 Z_{t-1} + \dots + \theta_q Z_{t-q}$$

3. $\{X_t\}$ is an $ARMA(p, q)$ process of order p and q if

$$X_t - \phi_1 X_{t-1} - \phi_2 X_{t-2} - \dots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2} + \dots + \theta_q Z_{t-q}$$

where $t = 0, \pm 1, \pm 2, \dots$ and $\phi_1, \phi_2, \dots, \phi_p$ and $\theta_1, \theta_2, \dots, \theta_q$ are constants. Furthermore, $\{Z_t\}$ is the error sequence such that, $Z_t \sim \text{white noise}(0, \sigma^2)$. That is, $\{Z_t\}$ is a sequence of uncorrelated random variables, each with zero mean and variance σ^2 .

In order to estimate the model and its parameters $\phi_1, \phi_2, \dots, \phi_p, \theta_1, \theta_2, \dots, \theta_q$, the orders of the parameters, p and q , must be determined. After transforming the data the sample ACF can be examined to get an idea of potential p and q values. Order selection is then based on finding the values of p and q that minimize the Akaike Information Criterion (AIC), which will be discussed later in the section. Once the order of p and q are determined, the parameters of the model can be estimated using the method of Maximum Gaussian Likelihood. Suppose that $\{X_t\}$ is a stationary time series with mean zero and autocovariance function $\kappa(i, j) = E(X_i X_j)$. Let $\mathbf{X}_n = (X_1, \dots, X_n)'$ and let Γ_n denote the covariance matrix such that $\Gamma_n = E(\mathbf{X}_n \mathbf{X}_n')$. Thus, the likelihood function of \mathbf{X}_n is

$$L(\Gamma_n) = \frac{1}{\sqrt{(2\pi)^n (\det \Gamma_n)}} \exp \left(-\frac{1}{2} \mathbf{X}_n' \Gamma_n^{-1} \mathbf{X}_n \right) \quad (1.1)$$

which using the Innovation Algorithm, can be reduced to

$$L(\Gamma_n) = \frac{1}{\sqrt{(2\pi)^n \nu_0 \cdots \nu_{n-1}}} \exp \left(-\frac{1}{2} \sum_{j=1}^n (X_j - \hat{X}_j)^2 / \nu_{j-1} \right) \quad (1.2)$$

where \hat{X}_j is the one-step prediction and $\nu_{j-1} = E(X_j - \hat{X}_j)^2$ is the mean squared prediction error.

The parameters are estimated by maximizing the likelihood function and the best model is selected based on minimizing the AIC. That is, selecting the values of p and q for the fitted model so as to minimize the AIC function,

$$\text{AIC} = -2 \ln L + \frac{2(p+q+1)n}{(n-p-q-2)}.$$

Satisfying the minimum AIC criterion provides a rational method for choosing between competing models, which can further be assessed based on the residuals (plots and tests of randomness) of the model. Details of assessing models based on their residuals will be discussed in Chapter 3.

1.3 Cross-Correlation Approach for Testing the Independence of Two Series

Much of the theory of univariate time series extends to the multivariate case and, in particular, bivariate time series, allowing for testing the independence of two stationary time series. Let $\mathbf{X}_t = (X_{t1}, X_{t2})'$ be a bivariate time series whose mean vector μ is the vector of sample means

$$\bar{\mathbf{X}}_n = \frac{1}{n} \sum_{t=1}^n \mathbf{X}_t.$$

A natural estimator of the covariance $\Gamma(h) = E[(\mathbf{X}_{t+h} - \mu)(\mathbf{X}_t - \mu)']$ is

$$\hat{\Gamma}(h) = \begin{cases} n^{-1} \sum_{t=1}^{n-h} (\mathbf{X}_{t+h} - \bar{\mathbf{X}}_n)(\mathbf{X}_t - \bar{\mathbf{X}}_n)' & \text{for } 0 \leq h \leq n-1, \\ \hat{\Gamma}'(-h) & \text{for } -n+1 \leq h < 0. \end{cases}$$

Writing $\hat{\gamma}_{ij}(h)$ for the (i, j) -component of $\hat{\Gamma}(h)$, $i, j = 1, 2, \dots$, the cross-correlations are estimated by,

$$\hat{\rho}_{ij}(h) = \hat{\gamma}_{ij}(h) (\hat{\gamma}_{ii}(0) \hat{\gamma}_{jj}(0))^{-1/2}.$$

Theorem 1: *Let $\{\mathbf{X}_t\}$ be the bivariate time series whose components are defined by*

$$X_{t1} = \sum_{k=-\infty}^{\infty} \alpha_k Z_{t-k,1}, \quad \{Z_{t1}\} \sim \text{IID}(0, \sigma_1^2),$$

and

$$X_{t2} = \sum_{k=-\infty}^{\infty} \alpha_k Z_{t-k,2}, \quad \{Z_{t2}\} \sim \text{IID}(0, \sigma_2^2),$$

where the two sequences $\{Z_{t1}\}$ and $\{Z_{t2}\}$ are independent, $\sum_k |\alpha_k| < \infty$ and $\sum_k |\beta_k| < \infty$. Then for all integers h and k with $h \neq k$, the random variables $n^{1/2} \hat{\rho}_{12}(h)$ and $n^{1/2} \hat{\rho}_{12}(k)$ are approximately bivariate normal with mean $\mathbf{0}$, variance $\sum_{j=-\infty}^{\infty} \rho_{11}(j) \rho_{22}(j)$ and covariance $\sum_{j=-\infty}^{\infty} \rho_{11}(j) \rho_{22}(j + k - h)$, for n large (Brockwell and Davis, 2002).

Theorem 1 is used to test the correlation between two time series and can provide some insight as to the relationship of the two series. However, since the large-sample distribution of a bivariate series depends on both $\rho_{11}(\cdot)$ and $\rho_{22}(\cdot)$, a test for independence of two series cannot be based solely on the estimated values of $\rho_{12}(h)$, $h = 0, \pm 1, \dots$, without taking into account the nature of the two component series. This can be corrected either by transforming each component series to white noise and then inspecting the cross-correlation of the two series or inspecting the cross-correlations after replacing the sequences $\{Z_{ti}\}$ by their residuals $\{\hat{W}_{ti}\}$ from fitting a maximum likelihood model to each component series.

Testing the hypothesis H_0 , that $\{X_{t1}\}$ and $\{X_{t2}\}$ are independent corresponds to testing the white noise series $\{Z_{t1}\}$ and $\{Z_{t2}\}$ for independence. By Theorem 1, the sample autocorrelations of $\{Z_{t1}\}$ and $\{Z_{t2}\}$ are independent and normally distributed with means 0 and variances n^{-1} , for large n , thus an approximate test for independence can be obtained by comparing the values of $|\hat{\rho}_{12}(h)|$ with $1.96n^{-1/2}$. Pre-whitening the series and/or finding appropriate models that fit the model is sometimes difficult, thus this approach is not always the best method for testing the independence of two time series. A more appropriate approach is to use structural time series with explanatory variables, which will be discussed in the following chapter.

Chapter 2

Structural Time Series

In this chapter a very useful and powerful technique for modeling a variety of time series models will be discussed. Here state space models are introduced along with the Kalman filter. Structural time series models incorporate the main observational features of most times series models such as trends and seasonal variations, thus detrending and deseasonalizing of the series is not required as in ARIMA models. The ideas of structural time series presented in this chapter will be based on Harvey, 1989.

2.1 State-Space Representation

The general state space form (SSF) applied to a multivariate time series $\{\mathbf{y}_t, \quad t = 1, 2, \dots\}$ containing N elements consists of two equations; the measurement equation (or observation equation) and the transition equation. The *measurement equation* is such that

$$\mathbf{y}_t = \mathbf{Z}_t \alpha_t + \mathbf{d}_t + \epsilon_t, \quad t = 1, 2, \dots, T \quad (2.1)$$

where \mathbf{Z}_t is an $N \times m$ matrix, \mathbf{d}_t is an $N \times 1$ vector, ϵ_t is an $N \times 1$ white noise (WN) vector such that $\epsilon_t \sim \text{WN}(0, \{\mathbf{H}_t\})$, T is the sample size of the series and α_t is known as the state vector. In general the elements of α_t are not observable, but are assumed to be generated by the *transition equation*,

$$\alpha_t = \mathbf{T}_t \alpha_{t-1} + \mathbf{c}_t + \mathbf{R}_t \eta_t, \quad t = 1, 2, \dots, T \quad (2.2)$$

where \mathbf{T}_t is an $m \times m$ matrix, \mathbf{c}_t is an $m \times 1$ vector, \mathbf{R}_t is an $m \times g$ matrix and \mathbf{Q}_t is a $g \times g$ matrix such that $\eta_t \sim \text{WN}(0, \{\mathbf{Q}_t\})$. Finally, the SSF is completed by two other assumptions:

1. the initial state vector, α_0 , has mean of \mathbf{a}_0 and a covariance matrix \mathbf{P}_0 , that is $E(\alpha_0) = \mathbf{a}_0$ and $\text{Var}(\alpha_0) = \mathbf{P}_0$.
2. the disturbances ϵ_t and η_t are uncorrelated with each other in all time periods and uncorrelated with the initial state, that is $E(\epsilon_t \eta_s') = 0$ for all $s, t = 1, \dots, T$ and $E(\epsilon_t \alpha_0') = 0$, $E(\eta_t \alpha_0') = 0$ for $t = 1, \dots, T$.

2.2 The Kalman Filter

In this section the concern is focused on finding the best linear estimates of the state vector α_t of the SSF defined by equations 2.1 and 2.2 in terms of the observations $\mathbf{y}_1, \mathbf{y}_2, \dots$, and the random vector \mathbf{y}_0 which in most cases is the constant vector $(1, 1, \dots, 1)'$, via the Kalman filter. Firstly, let \mathbf{a}_{t-1} denote the best linear mean-square predictors of α_{t-1} based on the observations $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_{t-1}$ (i.e. the best linear combination of $\mathbf{y}_0, \mathbf{y}_1, \dots, \mathbf{y}_{t-1}$ that minimizes the mean-squared error) and let \mathbf{P}_{t-1} denote the $m \times m$ covariance matrix of the estimation error, that is

$$\mathbf{P}_{t-1} = E[(\alpha_{t-1} - \mathbf{a}_{t-1})(\alpha_{t-1} - \mathbf{a}_{t-1})']. \quad (2.3)$$

Now, given \mathbf{a}_{t-1} and \mathbf{P}_{t-1} , the optimal estimator of α_t is given by

$$\mathbf{a}_{t|t-1} = \mathbf{T}_t \mathbf{a}_{t-1} + \mathbf{c}_t \quad (2.4)$$

while the covariance matrix of the estimation error is

$$\mathbf{P}_{t|t-1} = \mathbf{T}_t \mathbf{P}_{t-1} \mathbf{T}_t' + \mathbf{R}_t \mathbf{Q}_t \mathbf{R}_t', \quad t = 1, 2, \dots, T. \quad (2.5)$$

Equations 2.4 and 2.5 are known as the *prediction equations*, whereas once the new observation \mathbf{y}_t becomes available, the estimator of α_t , $\mathbf{a}_{t|t-1}$ can be updated by the following *updating equations*

$$\mathbf{a}_t = \mathbf{a}_{t|t-1} + \mathbf{P}_{t|t-1} \mathbf{Z}_t' \mathbf{F}_t^{-1} (\mathbf{y}_t - \mathbf{Z}_t \mathbf{a}_{t|t-1} - \mathbf{d}_t) \quad (2.6)$$

and

$$\mathbf{P}_t = \mathbf{P}_{t|t-1} - \mathbf{P}_{t|t-1} \mathbf{Z}_t' \mathbf{F}_t^{-1} \mathbf{Z}_t \mathbf{P}_{t|t-1} \quad (2.7)$$

where

$$\mathbf{F}_t = \mathbf{Z}_t \mathbf{P}_{t|t-1} \mathbf{Z}_t' + \mathbf{H}_t, \quad t = 1, 2, \dots, T \quad (2.8)$$

Taken together the prediction and updating equations, equations 2.4 - 2.8, make up the Kalman filter where it is assumed that the inverse of \mathbf{F}_t exists.

The starting values for the Kalman filter may be specified in terms of \mathbf{a}_0 and \mathbf{P}_0 . These initial values can be specified in two ways. First, if it is assumed that all the elements of α_0 are fixed then the Kalman filter can be initialized by specifying $\mathbf{a}_0 = \alpha_0$ and $\mathbf{P}_0 = \mathbf{0}$, where α_0 is a parameter to be estimated. However, when the transition equation is non-stationary, the unconditional distribution of the state vector is not defined. Thus, if no prior information is available, the initial distribution of α_0 must be specified in terms of a diffuse non-informative prior, resulting in the second type of initialization called *diffuse initialization*. Here, the Kalman filter can be initialized as $\mathbf{a}_0 = \mathbf{0}$ and $\mathbf{P}_0 = \kappa \mathbf{I}$ where κ is a positive scalar and the diffuse prior is obtained as $\kappa \rightarrow \infty$.

Given the the initial conditions \mathbf{a}_0 and \mathbf{P}_0 , of a non-diffuse initialization, the Kalman filter produces the optimal estimator of the state vector as each new observation becomes available. When all T observations are processed, the filter yields the optimal estimator of the current state vector, and/or the state vector in the next time period, based on the full information set. This estimator contains all the information needed to make optimal predictions of future values of both the state and the observations.

2.3 Estimation for State-Space Models

Consider the state-space model defined by equations (2.1) and (2.2) and suppose that the model is completely parameterized by the components of the vector Ψ . The likelihood function can be written as a conditional probability density function such that

$$L(\mathbf{y}; \Psi) = \prod_{t=1}^T f(\mathbf{y}_t | \mathbf{Y}_{t-1}) \quad (2.9)$$

where $f(\mathbf{y}_t|\mathbf{Y}_{t-1})$ denotes the distribution of \mathbf{y}_t conditional on the information set at time $t - 1$, that is $\mathbf{Y}_{t-1} = \{\mathbf{y}_{t-1}, \mathbf{y}_{t-2}, \dots, \mathbf{y}_1\}$. In time series analysis, the Gaussian likelihood is widely used whether the time series is truly Gaussian or not. Thus, the Gaussian likelihood function of the observations can be written as

$$\log L = -\frac{NT}{2} \log 2\pi - \frac{1}{2} \sum_{t=1}^T \log |\mathbf{F}_t| - \frac{1}{2} \sum_{t=1}^T \nu_t' \mathbf{F}_t^{-1} \nu_t \quad (2.10)$$

where

$$\nu_t = \mathbf{y}_t - \tilde{\mathbf{y}}_{t|t-1}, \quad t = 1, \dots, T \quad (2.11)$$

and

$$\tilde{\mathbf{y}}_{t|t-1} = \mathbf{Z}_t \mathbf{a}_{t|t-1} + \mathbf{d}_t \quad (2.12)$$

The prediction errors ν_t of equation (2.11) are known as *innovations*, since they represent the new information in the latest observation and $\tilde{\mathbf{y}}_{t|t-1}$ of equation (2.12) is the conditional mean of \mathbf{y}_t at time $t - 1$ and can be interpreted as the *minimum mean square estimator* (MMSE) of α_t . Thus, maximum likelihood estimates of the components of Ψ can be found by maximizing the likelihood function in equation (2.10).

Furthermore, let $\tilde{\Psi}$ denote the maximum likelihood estimator of the $n \times 1$ vector Ψ obtained by maximizing equation 2.10 and let the ij -th element of the information matrix $\mathbf{I}(\Psi)$, be defined as

$$-\mathbf{E} \left[\frac{\partial^2 \log L}{\partial \Psi_i \partial \Psi_j} \right] = -\mathbf{E} \left[\sum_{t=1}^T \frac{\partial^2 l_t}{\partial \Psi_i \partial \Psi_j} \right].$$

Suppose that $\mathbf{I}(\Psi)$, when divided by T , converges to a positive definite matrix, $\mathbf{IA}(\Psi)$. That is, $\mathbf{IA}(\Psi) = \lim_{T \rightarrow \infty} T^{-1} \mathbf{I}(\Psi)$. Subject to certain regularity conditions $\sqrt{T}(\tilde{\Psi} - \Psi)$ has a limiting multivariate normal distribution with mean vector zero and covariance matrix $\mathbf{IA}^{-1}(\Psi)$. Similarly, it can be stated that $\tilde{\Psi}$ is *asymptotically normal* with mean Ψ and covariance matrix $\text{Avar}(\tilde{\Psi}) = T^{-1} \mathbf{IA}^{-1}(\Psi)$.

2.4 Predictor Variables and Lagged Dependent Variables

When predictor variables, $\mathbf{x}_t'\beta$ are included in the model, such that $y_t = \mathbf{z}_t'\alpha_t + \mathbf{x}_t'\beta + \epsilon_t$, where $\beta = [\beta_1, \beta_2, \dots, \beta_n]'$ and $\mathbf{x}_t = [x_{1t}, x_{2t}, \dots, x_{nt}]'$, it is often the case that β is unknown, thus it is useful to incorporate it into the state vector, giving an augmented state vector

$$\alpha_t^\dagger = \begin{bmatrix} \alpha_t' & \beta_t' \end{bmatrix}. \quad (2.13)$$

Thus, when including the coefficients of the predictor variables into the model the augmented state vector, α_t^\dagger satisfies the SSF

$$\mathbf{y}_t = \begin{bmatrix} \mathbf{z}_t' & \mathbf{x}_t' \end{bmatrix} \alpha_t^\dagger + \epsilon_t, \quad t = 1, \dots, T \quad (2.14)$$

and

$$\alpha_t^\dagger = \begin{bmatrix} \alpha_t \\ \beta_t \end{bmatrix} = \begin{bmatrix} \mathbf{T} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \alpha_{t-1} \\ \beta_{t-1} \end{bmatrix} + \begin{bmatrix} \eta_t \\ \mathbf{0} \end{bmatrix}. \quad (2.15)$$

The lower part of the transition equation simply reflects the fact that $\beta = \beta_t$ is time-invariant. Including it in the state vector allows it to be estimated simultaneously with α_t . The parameters of the model are the disturbance variances and the regression coefficients. The disturbance variances, being elements of the system matrix are estimated by maximizing the likelihood, as described earlier, while the regression parameters get implicitly estimated during the state estimation.

When lags of the dependent variable are included in the model as $\mathbf{y}_t^p \text{prime} \phi$, such that $y_t = \mathbf{z}_t'\alpha_t + \mathbf{y}_t'\phi + \epsilon_t$, where $\mathbf{y}_t = [y_t, \dots, y_{t-r+1}]'$ and $\phi = [\phi_1, \dots, \phi_r]'$ (i.e. up to r dependent lags), y_t and its lagged values are included in the state vector, such that the SSF is

$$y_t = \begin{bmatrix} \mathbf{0}_m' & 1 & \mathbf{0}_{r-1}' \end{bmatrix} \alpha_t^* \quad (2.16)$$

$$\alpha_t^* = \begin{bmatrix} \alpha_t \\ \mathbf{y}_t \end{bmatrix} = \begin{bmatrix} \mathbf{T} & \mathbf{0} & \mathbf{0} \\ \mathbf{z}_t'\mathbf{T} & \phi' & \mathbf{0}' \\ \mathbf{0} & \mathbf{I}_{r-1} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \alpha_{t-1} \\ \mathbf{y}_{t-1} \end{bmatrix} + \begin{bmatrix} \mathbf{R}\eta_t \\ \mathbf{z}_t'\mathbf{R}\eta_t + \epsilon_t \\ \mathbf{0} \end{bmatrix} \quad (2.17)$$

where m is the number of elements in the state vector excluding the lagged dependent variables.

The parameters of this model are the disturbance variances and the lag coefficients $\phi_1, \phi_2, \dots, \phi_r$. Since these lag coefficients are not included in the state vector as the predictor coefficients explained in the previous paragraph, the lag coefficients along with the disturbance variances are estimated by maximizing the likelihood.

Chapter 3

Methodology

3.1 General Structural Time Series Model

Time series models incorporate the main features of most time series models, including trends and seasonal variations. The information presented in this chapter regarding the features of time series models comes from *The Unobserved Component Model Procedure* of the SAS, version 9.1, manual. The following equation represents the structural time series model used to identify various time series.

$$y_t = \mu_t + \gamma_t + \psi_t + \sum_{i=1}^p \phi_i y_{t-i} + \sum_{j=1}^m \beta_j x_{jt} + \epsilon_t \quad (3.1)$$

The component y_t is the dependent variable at time t and the components μ_t , γ_t and ψ_t model the trend, seasonal and cyclical components respectively and ϕ and β are the regression components. These different aspects of the time series are assumed to be statistically independent of each other and with the irregular component, ϵ_t . Below is a description of each of the components of the series.

The *trend* of the series consists of both the level (μ) and the slope (β) and can be described as:

$$\begin{aligned} \mu_t &= \mu_{t-1} + \beta_{t-1} + \eta_t & \eta_t &\sim i.i.d. \ N(0, \sigma_\eta^2) \\ \beta_t &= \beta_{t-1} + \xi_t & \xi_t &\sim i.i.d. \ N(0, \sigma_\xi^2). \end{aligned}$$

Some special cases arise here. If $\sigma_\xi^2 = 0$ then you obtain a model with fixed slope. If $\sigma_\eta^2 = 0$ then the resulting model usually has a smoother trend. If both, $\sigma_\xi^2 = \sigma_\eta^2 = 0$ the the resulting model is

the deterministic linear time trend: $\mu_t = \mu_0 + \beta_0 t$. Finally, if the trend remains roughly constant without any persistent upward or downward drift then no slope component exists resulting in the random walk model: $\mu_t = \mu_{t-1} + \eta_t$, $\eta_t \sim i.i.d. N(0, \sigma_\eta^2)$.

Seasonal fluctuations are common in time series data and arise because of the regular changes in seasons or some other periodic events. The *seasonal* component γ_t is modeled as a stochastic periodic pattern of an integer period s (season length) such that the sum $\sum_{i=0}^{s-1} \gamma_{t-i}$ is always zero in the mean. The most common type of model is called the *trigonometric* form of the seasonal component. Here γ_t is modeled as a sum of cycles of different frequencies and is given as,

$$\gamma_t = \sum_{j=1}^{\lfloor s/2 \rfloor} \gamma_{j,t} \quad (3.2)$$

where $\lfloor s/2 \rfloor$ equals $s/2$ if s is even and equals $(s-1)/2$ if it is odd. The frequencies here are $\lambda_j = 2\pi j/s$ and are specified by the following equations,

$$\begin{aligned} \gamma_{j,t} &= \gamma_{j,t-1} \cos \lambda_j + \gamma_{j,t-1}^* \sin \lambda_j + \omega_{j,t} \\ \gamma_{j,t}^* &= -\gamma_{j,t-1} \sin \lambda_j + \gamma_{j,t-1}^* \cos \lambda_j + \omega_{j,t}^* \end{aligned}$$

for $j = 1, \dots, \lfloor s/2 \rfloor$ where $\omega_{j,t}$ and $\omega_{j,t}^* \sim N(0, \sigma_\omega^2)$ and are assumed to be independent for fixed j . It is noted that when s is even, the component at $j = s/2$ collapses to

$$\gamma_{j,t} = \gamma_{j,t-1} \cos \lambda_j + \omega_{j,t}, \quad j = s/2 \quad (3.3)$$

Another form of the seasonal component is called the *dummy* variable form of the seasonal component and is described as

$$\sum_{i=0}^{s-1} \gamma_{t-i} = \omega_t, \quad \omega_t \sim i.i.d. N(0, \sigma_\omega^2) \quad (3.4)$$

In both the trigonometric and dummy forms of the seasonal component, if the disturbance variance $\sigma_\omega^2 = 0$ then both forms of the seasonal component reduce to a constant seasonal effect which is a deterministic function completely determined by its first $s-1$ values.

Another way of modeling the periodic pattern of a time series is by considering a *cycle* component. A deterministic cycle can be written as a mixture of sine and cosine waves such that

$$\psi_t = \alpha \cos(\lambda t) + \beta \sin(\lambda t). \quad (3.5)$$

where ψ_t has a frequency of λ , $0 < \lambda < \pi$, period $2\pi/\lambda$, amplitude $(\alpha^2 + \beta^2)^{1/2}$, phase $\tan^{-1}(\beta/\alpha)$ and t is measured on a continuous scale. However, it is more useful to consider a more general stochastic cycle that has a fixed period but time varying amplitude and phase by adding random noise and introducing a damping factor ρ to the model. The stochastic cycle considered here is described by the following recursive formulas,

$$\begin{aligned} \psi_t &= \rho(\psi_{t-1} \cos \lambda + \psi_{t-1}^* \sin \lambda) + \nu_t \\ \psi_t^* &= \rho(-\psi_{t-1} \sin \lambda + \psi_{t-1}^* \cos \lambda) + \nu_t^* \end{aligned}$$

where $0 \leq \rho \leq 1$ and the disturbances ν_t and ν_t^* are independent $N(0, \sigma_\nu^2)$ variables.

Introducing explanatory variables into a structural time series model is similar to the case of a standard regression model, and many of the concepts and modeling procedures associated with regression are relevant to the structural time series models that include explanatory variables. The regression terms $\sum_{i=1}^p \phi_i y_{t-i}$ and $\sum_{j=1}^m \beta_j x_{jt}$, where y_{t-i} and x_{jt} are lagged values of the dependent variable and other explanatory variables, respectively, and ϕ_i and β_j are the associated unknown parameters. $\sum_{i=1}^p \phi_i y_{t-i}$ considers the contribution of lagged values of the dependent variable to the model while $\sum_{j=1}^m \beta_j x_{jt}$ considers the contribution of other factors to the model. A variety of transformations including differences, lags and leads can be applied to the variables x_{jt} and included in the model.

3.2 Diagnostic Tests

In a well-specified model, the residuals should be approximately random. This section discusses various statistical tests of the standardized residuals $\tilde{\nu} = \nu_t / f_t^{1/2}$, where $f_t^{1/2}$ are the prediction standard errors, that are appropriate for assessing structural models.

Ljung-Box Test for Serial Correlation:

The residual sample autocorrelations are given by,

$$r_\nu(\tau) = \frac{\sum_{t=d+1+\tau}^T (\tilde{\nu}_t - \bar{\tilde{\nu}})(\tilde{\nu}_{t-\tau} - \bar{\tilde{\nu}})}{\sum_{t=d+1}^T (\tilde{\nu}_t - \bar{\tilde{\nu}})^2} \quad \tau = 1, 2, \dots \quad t = 1, 2, \dots, T \quad (3.6)$$

The test statistic of the first P residual autocorrelations is given by

$$Q^* = T^*(T^* + 2) \sum_{\tau=1}^P (T^* - \tau)^{-1} r_\nu^2(\tau) \quad (3.7)$$

where $T^* = T - c'$, such that d is the number of non-stationary elements of the state vector α_t . In a structural model, Q^* is asymptotically $\chi_{P-n^*}^2$, such that $n^* = n - 1$ where n is the number of hyperparameters in the model. A *hyperparameter* is a stochastic parameter estimated by the model. Thus, we reject the *i.i.d.* null hypothesis at a level of α if $Q^* > \chi_{1-\alpha, P-n^*}^2$.

Test for Heteroscedasticity:

Again a diagnostic test for heteroscedasticity can be constructed from the residuals. Here the test statistic to consider is given by

$$H(h) = \frac{\sum_{t=T-h+1}^T \tilde{\nu}_t^2}{\sum_{t=d+1}^T \tilde{\nu}_t^2} \quad (3.8)$$

where d is the same as above and h is the nearest integer to $T^*/3$. The $H(h)$ statistic can be tested against an $F(h, h)$ distribution, thus we reject the null hypothesis of homoscedasticity at level α if $H(h) > F(h, h)$.

Bowman-Shenton Test for Normality:

The *Bowman-Shenton* test for normality is based on the third and fourth moments of the residuals which are the basic measures of skewness and kurtosis of the residuals and are given respectively as

$$\sqrt{b_1} = \hat{\sigma}_*^{-3} \sum (\tilde{\nu}_t - \bar{\tilde{\nu}})^3 / T^* \quad (3.9)$$

and

$$b_2 = \hat{\sigma}_*^{-4} \sum (\tilde{\nu}_t - \bar{\tilde{\nu}})^4 / T^* \quad (3.10)$$

where $\hat{\sigma}_* = (T - d - 1)^{-1} \sum_{t=d+1}^T (\tilde{\nu}_t - \bar{\tilde{\nu}})^2$. For a normal distribution, equations (3.9) and (3.10) should be centered around zero and three respectively. The test statistic for normality is thus given by

$$N = (T^*/6)b_1 + (T^*/24)(b_2 - 3)^2 \quad (3.11)$$

Under the null hypothesis of normality, N is asymptotically χ_2^2 . Thus, we reject the null hypothesis if $N > \chi_2^2$.

Although the normality tests are standard diagnostic tests for model validity, the detection of non-normality in the residuals does not necessarily imply that the model is not good and a new model should be found. Non-normality in the residuals often arises due to outlier observations and structural breaks. These data irregularities often skew the results of the normality tests so that the null hypothesis is rejected in cases where it should not be rejected. Thus the inclusion of *intervention* or *dummy* variables into the model can correct for this. The idea of *dummy* and *intervention* variables is discussed in detail in the STAMP manual. An *outlier* which is an unusually large value of the *irregular* disturbances at a particular time in the model can be captured by including a dummy variable that takes on the value one at the time of the outlier and zero elsewhere as an explanatory variable in the model. On the other hand, a *structural break* in which the level of the series shifts up or down can be captured by a dummy variable which is zero before the event and one at and after the event as an explanatory variable in the model. The detection of irregular observations in the data can be determined by examining the *auxiliary residuals* (standardized smoothed estimates of the disturbances). Auxiliary residuals which have absolute value exceeding two are considered irregular observations and it is these observations that should be corrected for by the inclusion of dummy variables in the model.

The majority of the data modeling was done by SAS programming software, however when modeling data that requires the inclusion of dummy variables SAS encounters problems in terms of calculating residuals. Thus, for data sets that require the inclusion of dummy variables into the model, the statistical programming package STAMP is used which takes dummy variables into account properly. It was verified that both SAS and STAMP produce the same results (i.e. p-values) when dummy variables are not included in the model, thus STAMP was used with confidence when

dummy variables were included in the model.

3.3 Goodness of Fit

This section explains the goodness-of-fit statistics reported to measure how well the specified model fits the data. The various statistics of fit are computed using the *prediction errors* $y_t - \hat{y}_t$. In these formulae, n is the number of non-missing prediction errors (i.e. $T - d$). Recall that the sum of square errors, $SSE = \sum_{t=1}^n (y_t - \hat{y}_t)^2$ and the total sum of squares corrected for the mean, $SST = \sum_{t=1}^n (y_t - \bar{y})^2$, where \bar{y} is the series mean.

Mean Square Error (MSE):

The mean squared prediction error, MSE , is calculated from the one step ahead forecasts and is given as $MSE = SSE/n$.

R^2 :

R^2 is the conventional statistic calculated as $R^2 = 1 - (SSE/SST)$. The better the model fits the series, the closer the value of R^2 will be to unity. However, if the model fits the series poorly, the model error sum of squares SSE may be larger than SST and the R^2 statistic will be negative. Thus, a negative R^2 value indicates a poor model.

R_D^2 (Random Walk R^2):

A better measure for time series data is the Random Walk R^2 statistic R_D^2 obtained by replacing the observations by their first differences, that is

$$R_D^2 = 1 - \frac{SSE}{\sum_{t=2}^T (\Delta y_t - \overline{\Delta y})^2} \quad (3.12)$$

where $\overline{\Delta y}$ is the mean of the first differences. The model being used here is the simple random walk plus drift model,

$$y_t = y_{t-1} + \beta + \nu_t, \quad t = 2, \dots, T \quad (3.13)$$

This is a simple model in which the next period's forecast is taken to be the current observation plus the average increase over the sample period. Thus, $\sum_{t=2}^T (\Delta y_t - \overline{\Delta y})^2$ is simply the *SSE* for the model (3.13). Again, a model of good fit will have R_D^2 close to unity and a negative R_D^2 value indicates a poor model.

R_S^2 (Seasonally Adjusted R^2):

When a seasonal component is included in the model of the time series, a better indication of the fit of the model is given by the R_S^2 value. The adjustment for seasonal components can be done by simply including $s - 1$ seasonal dummies if there are s seasons to the model in (3.13). That is

$$y_t = y_{t-1} + \sum_{j=1}^s \gamma_j^* z_{tj} + \beta + \nu_t \quad t = 2, \dots, T \quad (3.14)$$

where z_{tj} 's are dummy variables taking the value in one season j and zero otherwise, and the γ_j^* 's are the unknown parameters. Thus, the goodness-of-fit statistic R_S^2 is

$$R_S^2 = 1 - \frac{SSE}{SSDSM} \quad (3.15)$$

where *SSDSM* is the sum of squares of first differences around the seasonal means. That is, *SSDSM* is simply the *SSE* for the model (3.14). Any model which has R_S^2 negative can be rejected, whereas R_S^2 positive but close to zero suggests that there is a marginal gain in model fit for a more complex model.

3.4 Model Selection

An essential preliminary step in model selection of a univariate time series is graphing the series. These graphs give insight as to the nature of the model. An unstable variance is often evident by examining the plot of the series, thus the log transformation of the series is initially taken to stabilize the variance of the series. It is from here that model selection begins. A general-to-specific approach is adopted for estimating the model and selecting the most appropriate model. This approach entails estimating a fully specified model with all the stochastic components and lagged values 1 and 2 of the dependent variable and then identifying significant components and 'testing-down'. That is, when there is no disturbance to a component in an estimated model (i.e. the hyperparameter is estimated to be approximately zero) a deterministic component can be included in the model instead and if this

deterministic component is not significantly different from zero then the respective component can be omitted all together and the model re-estimated. The most appropriate model is then selected based on the smallest AIC and the largest R_D^2 .

Once a final model is selected it is subjected to a series of diagnostic tests to assess its validity. If the model passes the series of tests then the model is considered valid. If the model is not valid based on the diagnostic tests, transformations of the dependent variable and/or the addition of dummy variables should be considered. Some of the transformations of the dependent variable to be considered are:

1. $\sqrt{y_t}$
2. $\sqrt{\sqrt{y_t}}$
3. $\sqrt{1 + y_t}$
4. $\frac{(\sqrt{y_t} + \sqrt{1 + y_t})}{2}$
5. $\frac{1}{y_t}$

Once a valid model is found, the significance and estimation of any explanatory variable (other than lagged values because they are considered in the initial model) can be assessed by adding it to the existing valid model and re-estimating the model. If the explanatory variable is significant then the new model with the explanatory variable is re-estimated and its residual tests re-assessed to ensure the new model is valid. If the new model produces residual tests that suggest the new model is not good, the explanatory variable should be subjected to some transformation similar to the ones mentioned above or further dummy variables added to the new model so that the residual tests suggest that the model is good. Once a valid model is found, model selection and identification are complete.

Chapter 4

Modeling Example

In this chapter a detailed example of the modeling procedure used for the time series data will be outlined. In this example, the asthma data for patients under two years (ASU2) along with viral data (RSV and FLU AB) is used. Initially, the ASU2 data is plotted with each of the viral data, RSV and FLU AB, separately and the resulting graphs can be found in Figure 4.1 and Figure 4.2 respectively.

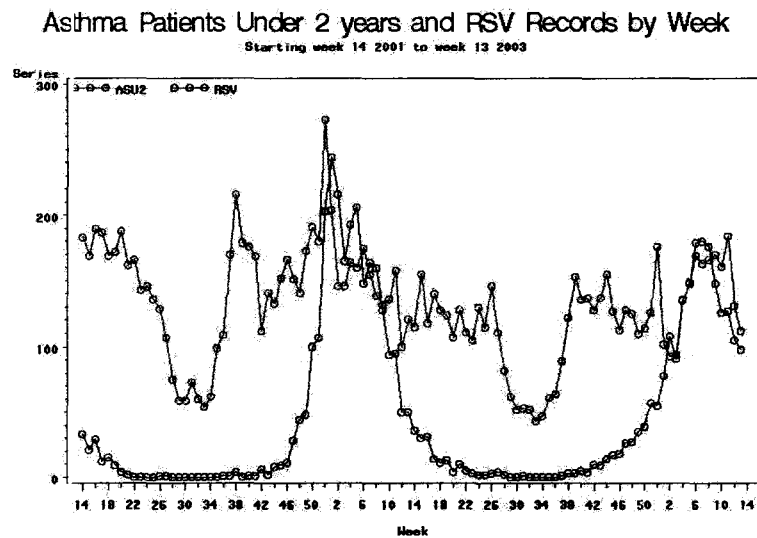


Figure 4.1: *Plot of asthma patients under 2 years and RSV positive tests (over mean)*

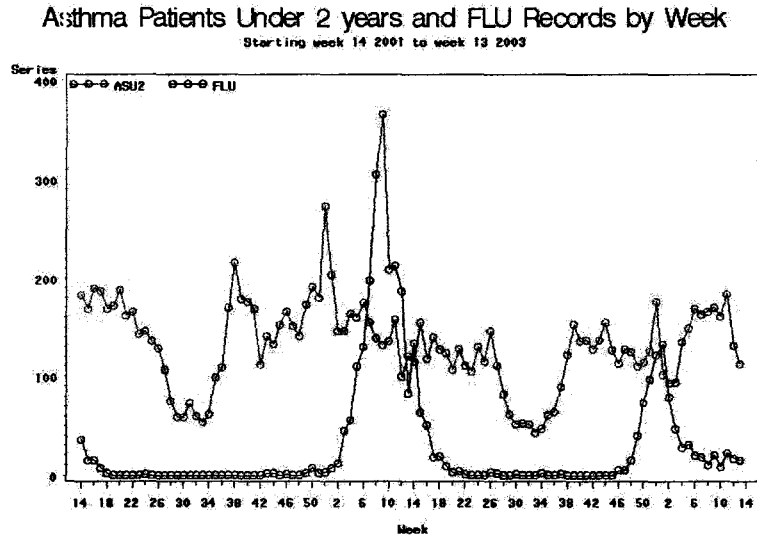


Figure 4.2: *Plot of asthma patients under 2 years and FLU AB positive tests (over mean)*

To get some insight as to the relationship of ASU2 and RSV and FLU AB, the cross-correlations of ASU2 with RSV and FLU AB were plotted individually, following the method outlined in section 1.3. Below are the plots of the cross-correlations from the differenced data of the log transformed ASU2 data and the differenced RSV and FLU AB data. Transformations of the ASU2 and RSV and FLU AB data were considered in order to obtain stationary data, since the cross-correlation method described in section 1.3 is best suited for stationary data.

From Figure 4.3, it is evident that the greatest correlation occurs at lag 0 (i.e. no time difference). At lag 0 there exists a positive correlation between the ASU2 and RSV positive tests. At lag 1 there exists a small negative correlation between ASU2 and RSV a week behind. At lag 2 there exists a small positive correlation between ASU2 and RSV two weeks behind. The remainder of the graph can be interpreted similarly.

From Figure 4.4, it is evident that a positive correlation occurs at lag 0. At lag 1 there exists a small positive correlation between RSV positive tests and ASU2 a week behind (i.e. positive correlation between ASU2 and RSV a week ahead). At lag 2 there exists a small positive correlation between RSV positive tests and ASU2 two weeks behind (i.e. positive correlation between ASU2 and RSV two weeks ahead).

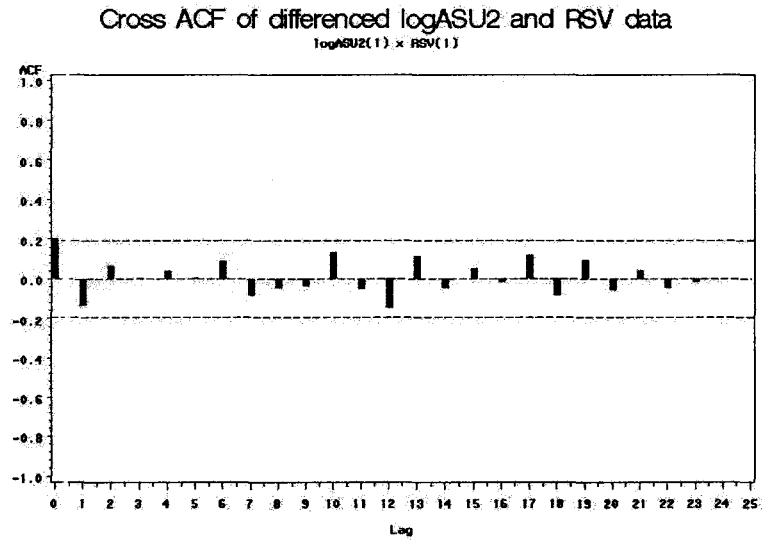


Figure 4.3: *Cross Correlation Plot of transformed ASU2 and RSV data*

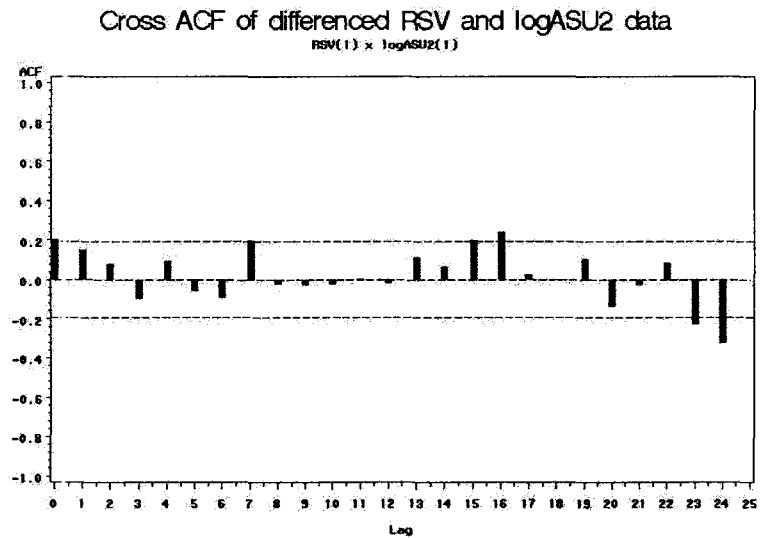


Figure 4.4: *Cross Correlation Plot of transformed RSV and ASU2 data*

The remainder of the graph can be interpreted similarly. It can be noted that Figure 4.3 pertains to the lagged values of the RSV data, while Figure 4.4 pertains to the lead values of the RSV data.

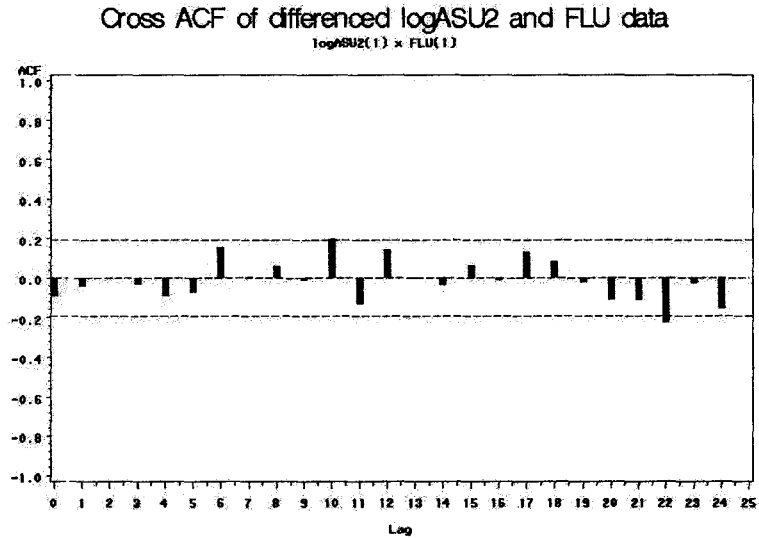


Figure 4.5: *Cross Correlation Plot of transformed ASU2 and FLU AB data*

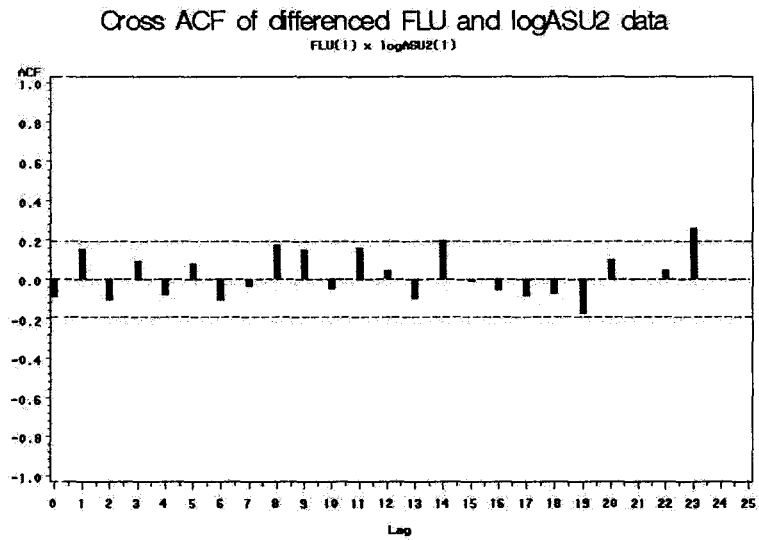


Figure 4.6: *Cross Correlation Plot of transformed FLU AB and ASU2 data*

The plots of the cross correlations between ASU2 and FLU AB can be interpreted in the same way the previous plots (Figure 4.3 and Figure 4.4) were interpreted for ASU2 and RSV. Figure 4.5 pertains to lagged values of FLU AB and Figure 4.6 pertains to lead values of FLU AB.

Next, an appropriate model is found for the ASU2 data by using a general to specific approach, starting with the most broad model, including all stochastic components and dependent lags 1 and 2, and testing down. First, the broad model was applied to both the untransformed ASU2 data and the log transformed ASU2 data to determine if the log transformation indeed improves the model by stabilizing the variance. The fit statistics (including A.I.C., MSE, R^2 and R_D^2) of the ASU2 model and the log(ASU2) model are found in Table 4.1 and Table 4.2 respectively.

Table 4.1: Fit Statistics of ASU2 Model

| A.I.C. | MSE | R^2 | R_D^2 |
|-----------|-----------|---------|---------|
| 960.33607 | 648.99367 | 0.62644 | 0.34117 |

Table 4.2: Fit Statistics of log(ASU2) Model

| A.I.C. | MSE | R^2 | R_D^2 |
|-----------|---------|---------|---------|
| -14.27072 | 0.03839 | 0.72684 | 0.87177 |

By examining Table 4.1 and Table 4.2, it is noted that the log(ASU2) model is indeed better than the untransformed ASU2 model. Thus, model selection continues based on the log transformed ASU2 data. The log transformed ASU2 model, including all stochastic components (level, slope, irregular and cycle) will be denoted as Model 1. The corresponding parameter output for Model 1 is given below in Table 4.3 and the significance analysis of the components in Model 1 is given in Table 4.4.

Table 4.3: Parameter Estimates of Model 1

| Component | Parameter | Estimate | Std. Error | t Value | Pr > $ t $ |
|-----------|----------------|-------------|------------|-----------|------------|
| Irregular | Error Variance | 7.62492E-10 | 5.02301E-6 | 0.00 | 0.9999 |
| Level | Error Variance | 0.03608 | 0.0051555 | 7.00 | <.0001 |
| Slope | Error Variance | 2.6017E-13 | 3.92139E-9 | 0.00 | 0.9999 |
| Cycle | Damping Factor | 1.00000 | 0.0001306 | 7658.38 | <.0001 |
| Cycle | Period | 6.46410 | 0.08750 | 73.87 | <.0001 |
| Cycle | Error Variance | 1.145245E-8 | 1.50364E-8 | 0.76 | 0.4463 |
| DepLag | Phi1 | -0.02090 | 0.09771 | -0.21 | 0.8306 |
| DepLag | Phi2 | 0.09243 | 0.10199 | 0.91 | 0.3648 |

From section 2.3, the estimates of Ψ , estimated from the likelihood function, equation 2.10, follow an asymptotic normal distribution with mean and covariance matrix as described in section 2.3. Since the observed standard error is used for evaluating the statistic, the estimates follow a t distribution.

Table 4.4: Significance Analysis of Components of Model 1

| Component | DF | Chi-Square | Pr > Chi-Square |
|-----------|----|------------|-----------------|
| Irregular | 1 | 0.00 | 0.999 |
| Level | 1 | 29176.7 | <0.0001 |
| Slope | 1 | 0.11 | 0.7387 |
| Cycle | 2 | 6.66 | 0.0359 |

Table 4.4 tests the validity of any restrictions placed on the estimated parameters of the model (i.e. Ψ). Under the null hypothesis, H_0 , the maximum likelihood (ML) estimator Ψ is restricted and is denoted by $\tilde{\Psi}_0$. The restricted ML estimator can be contrasted with the unrestricted estimator, $\tilde{\Psi}$. If the maximized likelihood function under H_0 , $L(\tilde{\Psi}_0)$, is much smaller than the unrestricted maximized likelihood, $L(\tilde{\Psi})$, there is evidence against the null hypothesis. This is the idea behind the likelihood ratio test, where the likelihood ratio is, $\lambda = L(\tilde{\Psi}_0)/L(\tilde{\Psi})$. Furthermore, the likelihood ratio statistic, $LR := -2\log\lambda$ is asymptotically distributed as χ_m^2 under H_0 , where m is the difference in the number of parameters to be estimated between the restricted and unrestricted models. Thus, the results of Table 4.4 are based on this idea.

From Table 4.3 and 4.4 it is evident, by examining the p-values that, the irregular and slope components are not significant and, by examining Table 4.3, that both dependent lags (1 and 2) are not significant components in the model. Thus, one proceeds by removing these components and re-evaluating the model. The new model that only consists of the level and cycle components will be denoted as Model 2. Table 4.5 provides the fit statistics of the Model 2, while Table 4.6 and Table 4.7 provide the parameter estimates and significance analysis of components of Model 2 respectively.

Table 4.5: Fit Statistics of Model 2

| A.I.C. | MSE | R^2 | R_D^2 |
|-----------|---------|---------|---------|
| -40.79207 | 0.03500 | 0.75182 | 0.87971 |

Table 4.6: Parameter Estimates Model 2

| Component | Parameter | Estimate | Std. Error | t Value | Pr > $ t $ |
|-----------|----------------|-------------|------------|-----------|------------|
| Level | Error Variance | 3.970704E-9 | 4.40098E-6 | 0.00 | 0.9993 |
| Cycle | Damping Factor | 0.88518 | 0.03683 | 24.04 | <.0001 |
| Cycle | Period | 34.80643 | 11.94365 | 2.91 | 0.0036 |
| Cycle | Error Variance | 0.03145 | 0.0099566 | 3.16 | 0.0016 |

Table 4.7: Significance Analysis of Components of Model 2

| Component | DF | Chi-Square | Pr > Chi-Square |
|-----------|----|------------|-----------------|
| Level | 1 | 3219.04 | <0.0001 |
| Cycle | 2 | 2.00 | 0.3675 |

From Table 4.7, it appears that the level component is significant, however from Table 4.6 it appears that the level error variance is not significant in the model. Although, the cycle component in Table 4.7 appears not to be significant in the model, all of its components are significant in Table 4.6, thus it should remain in the model. Therefore, the model is once again re-estimated by removing the level error variance, by setting it equal to zero in the model. The new model that consists of a cycle component and level component with level error variance set equal to zero will be denoted by Model 3. Table 4.8 provides the fit statistics of the Model 3, while Table 4.9 and Table 4.10 provide the parameter estimates and significance analysis of components of Model 3 respectively.

Table 4.8: Fit Statistics of Model 3

| A.I.C. | MSE | R^2 | R_D^2 |
|-----------|---------|---------|---------|
| -42.79207 | 0.03500 | 0.75182 | 0.87971 |

Table 4.9: Parameter Estimates Model 3

| Component | Parameter | Estimate | Std. Error | t Value | Pr > $ t $ |
|-----------|----------------|----------|------------|-----------|------------|
| Cycle | Damping Factor | 0.88518 | 0.03683 | 24.04 | <.0001 |
| Cycle | Period | 34.80643 | 11.94365 | 2.91 | 0.0036 |
| Cycle | Error Variance | 0.03145 | 0.0099566 | 3.16 | 0.0016 |

Table 4.10: Significance Analysis of Components of Model 2

| Component | DF | Chi-Square | Pr > Chi-Square |
|-----------|----|------------|-----------------|
| Level | 1 | 3219.04 | <0.0001 |
| Cycle | 2 | 2.00 | 0.3675 |

In order for the structural time series approach via Kalman filters to work, a model must contain at least one stochastic component, thus the model is not tested down further and should not be as Table 4.9 shows that the error variance of the cycle component is significant. Furthermore, by examining the fit statistics of the Models 1, 2 and 3 by examining Tables 4.2, 4.5 and 4.8 respectively, it is observed that Model 3, the simplest model has the smallest AIC value with the largest R_D^2 value, thus the appropriate model for the asthma under 2 years data is Model 3, level component (with level variance set to zero) and cycle component. Thus, model 3 is the most appropriate model for the ASU2 data.

Now that the best model has been selected for the ASU2 data, this model must now be verified by examining the residual plots and tests as discussed in section 3.2.

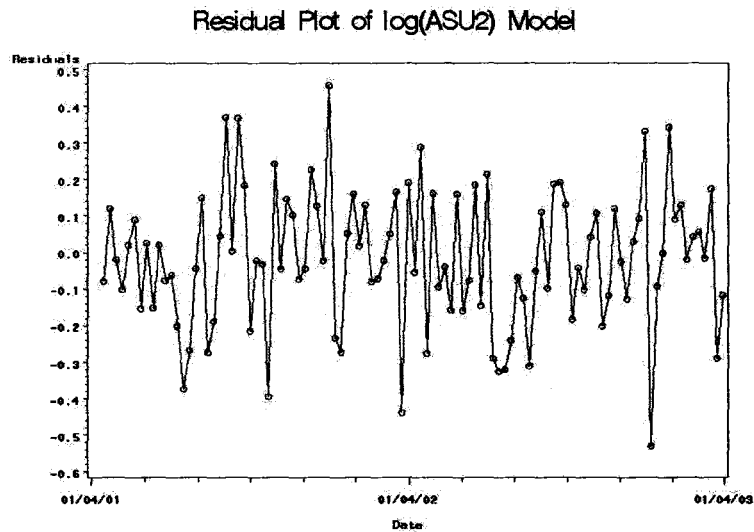


Figure 4.7: *Residual Plot of ASU2 model (Model 3)*

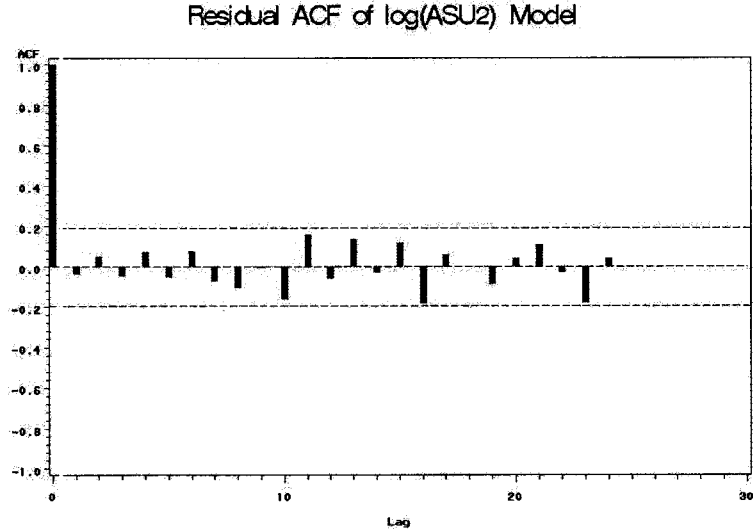


Figure 4.8: *ACF of the residuals of the ASU2 model (Model 3)*

Table 4.11: Ljung-Box Residual Test

| P | Chi-Square | DF | P-value |
|-----|------------|----|---------|
| 6 | 2.27 | 6 | 0.8937 |
| 12 | 10.68 | 12 | 0.5562 |
| 18 | 19.33 | 18 | 0.3716 |
| 24 | 26.70 | 24 | 0.3188 |

Table 4.12: Other Residual Tests

| Test | Statistic | P-value |
|--------------------|-----------|---------|
| Heteroscedasticity | 1.0040073 | 0.4954 |
| Bowman-Shenton | 0.1122158 | 0.9454 |

The residual plots for the ASU2 model (Figure 4.7 and Figure 4.8) appear reasonable and the residual tests (Table 4.11 and Table 4.12) do not give evidence against the model, therefore this model is an appropriate model for the ASU2 data. If the residual tests appeared to be unreasonable, it is at this point that dummy variables are introduced into the most model as explanatory variables in order to make the residuals appropriate.

A valid model has been found for the ASU2 data, thus testing the independence between ASU2 (dependent variable) and the viral infections (independent variables) RSV and FLU AB can proceed. In order to test for independence of ASU2 and the viral infections, the untransformed viral data

along with seven transformations of the viral data were considered. The seven transformations included were; differenced(1) (dif1), lag 1 (lag1), lag 2 (lag2), lag 1 of differenced(1) (lagdif), lag 2 of differenced(1) (lag2dif), lead 1 (lead1), lead 2 (lead2). The eight tests for each viral infection were considered in order to account for a variety of possibilities of the independent variable. Table 4.13 and Table 4.14 show the results of the eight tests for the RSV and FLU AB data respectively.

Table 4.13: RSV Predictors

| Component | Estimate | P-value | AIC | MSE | R^2 | R_D^2 | Adj. P-value |
|------------|------------|-----------|-----------|---------|---------|---------|--------------|
| RSV | 0.00241 | 0.0070 ** | -35.08301 | 0.03442 | 0.75666 | 0.88738 | 0.0550 |
| dif1RSV | 0.00204 | 0.0067 ** | -32.60532 | 0.03554 | 0.74788 | 0.88423 | 0.0550 |
| lag1RSV | -0.00122 | 0.2282 | -28.81985 | 0.03823 | 0.72878 | 0.87546 | 1.8259 |
| lag2RSV | 0.00125 | 0.1803 | -27.79605 | 0.04242 | 0.69822 | 0.85834 | 1.4428 |
| lagdifRSV | -0.00125 | 0.1072 | -26.65914 | 0.03623 | 0.74222 | 0.87899 | 0.8574 |
| lag2difRSV | 0.00040520 | 0.6111 | -23.03748 | 0.04088 | 0.70988 | 0.86565 | 4.8884 |
| RSVlead1 | 0.00194 | 0.0341 * | -31.48521 | 0.03646 | 0.74462 | 0.88172 | 0.2070 |
| RSVlead2 | 0.00150 | 0.1060 | -30.59984 | 0.03554 | 0.75345 | 0.88528 | 0.8477 |

Table 4.14: FLU AB Predictors

| Component | Estimate | P-value | AIC | MSE | R^2 | R_D^2 | Adj. P-value |
|------------|-------------|----------|-----------|---------|---------|---------|--------------|
| FluABPos | -0.00045651 | 0.5139 | -29.25170 | 0.03608 | 0.74491 | 0.88194 | 4.1111 |
| dif1FLU | -0.00022439 | 0.6575 | -25.77415 | 0.03631 | 0.74236 | 0.88169 | 5.2602 |
| lag1FLU | -0.00001176 | 0.9866 | -27.92046 | 0.03707 | 0.73701 | 0.87924 | 7.8930 |
| lag2FLU | 0.00027439 | 0.6967 | -26.62657 | 0.03786 | 0.73064 | 0.87355 | 5.5734 |
| lagdifFLU | -0.00013455 | 0.7915 | -24.21244 | 0.03869 | 0.72475 | 0.87079 | 6.3316 |
| lag2difFLU | 0.00011885 | 0.8165 | -22.84895 | 0.04013 | 0.71513 | 0.86808 | 6.5316 |
| FLUlead1 | 0.00151 | 0.0279 * | -32.01457 | 0.03795 | 0.73324 | 0.87741 | 0.2231 |
| FLUlead2 | -0.00066647 | 0.3391 | -29.54500 | 0.03867 | 0.73172 | 0.87517 | 2.7126 |

* : significant at a 5% level of significance

** : significant at a 1% level of significance

*** : significant at a 0.1% level of significance

It should be noted that the column “Adj. P-value” is an adjusted p-value column. The p-values obtained through the software were adjusted for by the *Bonferroni Adjustment* in order to keep the overall experiment rate to an α -level of 0.05. The α -level is the probability of making a type I error (i.e. error of incorrectly determining a factor to be significant when it is not significant). In the case of more than one statistical test, the chance of finding at least one test statistically significant due to chance in the total experiment, and hence incorrectly declare a significant effect, increases. Thus, in eight tests the chance of finding at least one relationship significant due to chance

fluctuation, assuming independence, equals 0.125, or one in eight. Using the Bonferroni method the α -level of each individual test is adjusted downwards to ensure that the overall experimentwise risk for a number of tests remains 0.05. Thus, the α -level for eight tests would be adjusted downward by dividing 0.05 by eight, resulting in an α -level of 0.00625. Equivalently, multiplying the p-value obtained by the test by eight and testing the resulting adjusted p-value against the α -level 0.05 is the exact same adjustment for eight tests as the previous statement. This latter method of adjustment was used for testing of independence in the project. The Bonferroni Adjustment ensures that if more than one test is done the risk of finding a difference or effect incorrectly significant continues to be less than 0.05. It should also be noted that simulations were also done to adjust for the p-values in the case that the Bonferroni adjustment method was too conservative. However, the simulations gave adjustment factors close to eight, thus the Bonferroni adjustment method was used.

By examining the adjusted p-values in Table 4.13 and Table 4.14, neither RSV nor FLU AB are significant predictors for the ASU2 data. In a case where a significant predictor existed, the residual plots and residual tests would be examined to verify the model.

Chapter 5

Results and Discussion

In this chapter the results of the 18 analyses done for this project are presented and discussed. The models are presented in table format and the components of the model along with the predictors are included. Components included in the model will be identified with a “x” are assumed to be stochastic if not otherwise indicated (i.e. $\text{var} = 0$) and components not included in the model are identified with a “-”. Furthermore, significant predictors are indicated by the test that was significant and the resulting adjusted p-value given in brackets, otherwise non-significant predictors are represented by “N.S.”. It should finally be noted that all the models are for the log transformed dependent variables.

5.1 Asthma Age Groups

Table 5.1: Asthma Age Group Models with Viral Infections (RSV FLU AB) as Predictors

| Age (years) | level | slope | cycle | irregular | lags | dummy variables (year/week) | RSV | FLU |
|-------------|-----------|-------|-------|-----------|-------|-----------------------------|------|------|
| Under 2 | x (var=0) | – | x | – | – | – | N.S. | N.S. |
| 2-4 | x (var=0) | – | x | – | lag 1 | – | N.S. | N.S. |
| 5-15 | x (var=0) | – | x | – | | 2002/1* 2003/1* | N.S. | N.S. |
| 16-49 | x (var=0) | – | x | x | lag 1 | 2001/52 2002/52 | N.S. | N.S. |
| Over 50 | – | – | x | – | lag 1 | – | N.S. | N.S. |

* : dummy variables are for structural break in the level component

By examining Table 5.1 it is evident that the viral infections RSV and FLU AB are not significant predictors of asthma hospitalizations in children under two years. Looking back at Figures 4.1 and 4.2, which plot ASU2 together with RSV and ASU2 together with FLU AB respectively, it is noted that the results shown in Table 15 correspond to the graphical display in these two plots as a direct relationship or correspondence between ASU2 and the viral infection is not evident. That is, by examining Figure 4.1 and Figure 4.2, the peaks and troughs of ASU2 do not graphically correspond with the peaks and troughs of the viral infection, thus confirming the result that RSV and FLU AB are not significant predictors of ASU2. Furthermore, by examining the other age groups of asthma, it is noted that RSV and FLU AB are not significant predictors of these other asthma age groups. Thus, no relationship was found between RSV and FLU AB and asthma hospitalizations in any age group.

5.2 RTI Age Groups

Table 5.2: RTI Age Group Models with Viral Infections (RSV FLU AB) as Predictors

| Age (years) | level | slope | cycle | irregular | lags | dummy variables (year/week) | RSV | FLU |
|-------------|-----------|-------|-------|-----------|----------------|---|------------------|-------------------|
| Under 2 | x (var=0) | – | x | x | lag 1 | 2001/52 2002/52 | dif1 (0.0032) | N.S. |
| 2-4 | x | – | x | – | – | 2001/52 2002/51 2002/52 | N.S. | N.S. |
| 5-15 | x | – | x | x | – | 2001/52 2002/51 2002/52 | N.S. | lead1 (0.0016) |
| 16-49 | – | – | x | x | lag 1 | 2001/52 2002/2 2002/14 2002/52 2003/1 2003/2 | N.S. | N.S. |
| Over 50 | x | – | – | x | lag 1 lag 2 | 2001/52 2002/52 | N.S. | N.S. |

RSV is strongly associated with RTI hospitalizations in children under 2 years and FLU AB is strongly associated with RTI hospitalizations in school aged children 5-15 years. By examining the transformations of the significant relationships, it is evident that RTI is related to the difference between consecutive RSV positive tests in children under 2 years, whereas RTI is related to FLU AB one week ahead in patients 5-15 years. In the other RTI age groups (2-4, 16-49 and over 50), RSV and FLU AB have no apparent association with RTI.

5.3 COPD Over 50

Table 5.3: COPD Over 50 Model with Viral Infections (RSV FLU AB) as Predictors

| Age (years) | level | slope | cycle | irregular | lags | dummy variables (year/week) | RSV | FLU |
|-------------|-------|-------|-------|-----------|----------------|-----------------------------|------|------|
| Over 50 | x | — | — | x | lag 1 lag 2 | 2001/52 2002/52 | N.S. | N.S. |

COPD hospitalization in adults over 50 years show no association with RSV and FLU AB.

5.4 Respiratory Disease Analysis By Age Group

In this section the respiratory diseases were tested for independence by age group. Table 5.4 gives the results of these analyses. The dependent variable is indicated along with the independent variable of the model. The initial model for the dependent variable is given in either Table 5.1, Table 5.2 or Table 5.3 excluding the viral infection predictors, thus the model components will not be displayed in Table 5.4 and only the significance of the independent variable will be noted.

Table 5.4: Respiratory Disease Analysis By Age Group

| Age (years) | Dependent variable | Independent variable | Significance |
|-------------|--------------------|----------------------|---------------------|
| Under 2 | Asthma | RTI | dif1 (2.0491626E-9) |
| 2-4 | RTI | Asthma | AS24 (0.008) |
| 5-15 | Asthma | RTI | lagdif (0.0104) |
| 16-49 | Asthma | RTI | lag2 (0.0128) |
| Over 50 | Asthma | RTI | dif1 (0.0000) |
| Over 50 | COPD | asthma | ASO50 (0.0000) |
| Over 50 | COPD | RTI | RTIO50 (0.0000) |

By examining Table 5.4, it is noted that the respiratory diseases are all strongly associated. Asthma has a strong relationship with RTI in all age groups and these relationships are based on different transformations of the RTI data across the age groups. However, COPD is strongly associated with the exact data (no transformations) of both asthma and RTI. As an example, by examining Figure 5.1 and Figure 5.2 below, which display the plots of COPD and asthma over 50 and COPD and RTI over 50 respectively, the strong association of respiratory diseases in adults over 50 is clearly evident as the plots are overlapping and follow very similar paths.

COPD Patients over 50 years and Asthma Patients over 50 years
Starting week 14 2001 to week 13 2003

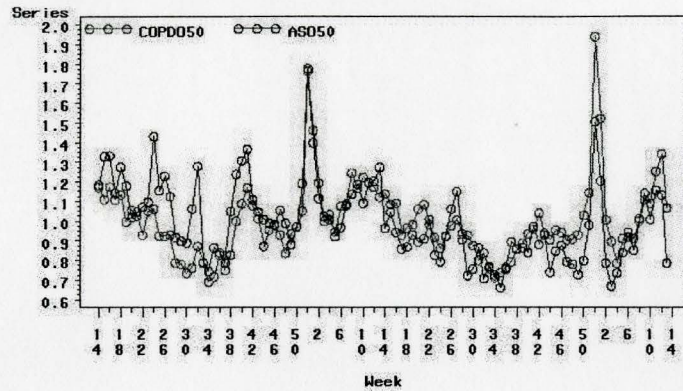


Figure 5.1: *COPD and Asthma Patients Over 50 (over mean)*

COPD Patients over 50 years and RTI Patients over 50 years
Starting week 14 2001 to week 13 2003

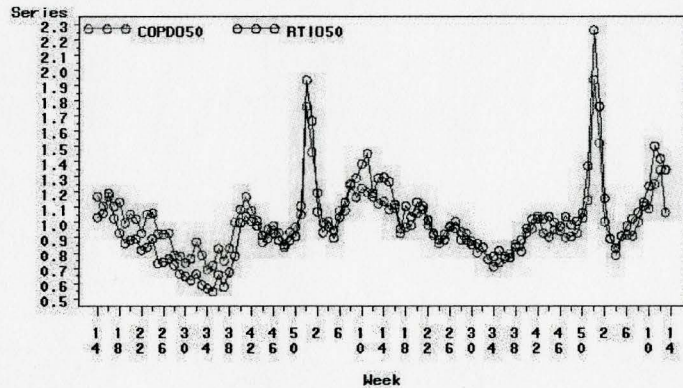


Figure 5.2: *COPD and RTI Patients Over 50 (over mean)*

5.5 Discussion

It is believed, yet not confirmed, that the influenza virus is often related to respiratory disease in young children and adults over 50 years and can have a severe effect on them. However, this relationship was not evident in the results presented in this project. Influenza is a virus with multiple strains and in any given year individuals may be effected by any of a number of these strains. Thus,

the influenza virus is unpredictable and the strain types and associated severity vary from year to year, with varying effects on respiratory disease patients yearly. Likewise, the capacity for influenza strains to exacerbations in COPD varies continually in an unknown fashion. Furthermore, with the availability of the influenza vaccination, the response of asthmatic, RTI and COPD patients to the influenza virus may also be affected, explaining the results presented. Thus, in a single year, it is perfectly feasible that no relationship may be found between the influenza virus and suspected respiratory disease children and adults. By the same token, the particular influenza strains for the virus in that year may have had a more severe affect on school-aged children, 5-15 years. The fact that children are exposed to other school-aged children throughout the year in a school setting makes them more susceptible to contracting and spreading the influenza virus and resulting in a respiratory tract infection. Thus, again it is feasible that in any given year a relationship between influenza and RTI in school-aged children 5-15 years exists.

On the other hand, while RSV is a single virus, its antigenic profile changes, permitting multiple or serial infections. It is possible then that in any given year the pathogenicity of RSV may vary itself or be influenced by cofactors, thus again it is possible that in a single year no effect maybe found with respiratory diseases over the various age groups (except children under 2 years) even though an effect may be found over multiple years. However, a significant relationship between RSV and RTI for children under 2 was found through the analysis. This relationship is considered common knowledge to professionals in the field of respiratory health as RSV is the number one cause of RTI in infants. Thus, the result found in the analysis confirms this idea.

Finally, the strong relationship between the respiratory diseases within each age group seems very reasonable. For children under 5 years (under 2 years, 2-4 years), asthma and RTI are difficult to distinguish between since their respiratory system is not fully developed yet. Furthermore, when these children are hospitalized, the main concern is to fix the problem, thus misdiagnosing and overlapping of diagnoses can occur. Furthermore, for the other age groups, when an individual with asthma or COPD suffers and is hospitalized with RTI, often the effects of RTI cause their asthma and COPD exacerbations to be increased, resulting in hospitalizations. Similarly, for adults over 50, either increased asthma or COPD exacerbations causes the other to be increased, resulting in a significant relationship.

Chapter 6

Conclusions and Future Work

The results have demonstrated successful application of structural time series modeling to the data and have successfully demonstrated that respiratory diseases over different age groups respond differently to different viral infections. Thus, the relationship between respiratory diseases and viral infections are unpredictable.

However, some ideas are suggested for the unpredictable results. First and foremost, the research and analysis performed for this project was based on data from week 14 of the year 2001 to week 13 of the year 2003 inclusive. By examining the plots of this data (Figure 4.1 and Figure 4.2), it is noted that these data include only one full peak of the viral infections and part of another peak. Thus, although the results of the analysis are valid for this data set, it may not be appropriate and effective to make these results conclusive for the population, since the second partial peak may be skewing the results. Thus, future studies should apply the same methodology to multiple years of data to determine more conclusive results. Furthermore, although the data are collected with caution, the timing of the reported viral infection positive tests are not as accurate and are, in fact, reported later than the week they were actually taken. Although this reporting discrepancy was accounted for by adjusting all the dates by a set period based on the average of the reporting time, this may also affect the results slightly.

Currently, I am doing work with the Firestone Institute for Respiratory Health at St. Joseph's Hospital, extending my research and analysis to an eight-year data set that includes seven peaks of the viral infections. The eight-year data are being examined and analysed as a complete data set and the seven individual peaks are being analysed separately for further verification, using the same

theory and methods applied throughout this project. I am also including simulations to find the appropriate p-values for the tests when the sample size is small. The larger data set will provide a more insightful look to the true relationship between respiratory diseases and viral infections of different age groups.

Appendix

Appendix A: SAS Program

This appendix includes all the SAS programming code used for the analyses done using SAS in this project.

PLOTTING ORIGINAL DATA PROGRAM:

```
/* Calls the comma separated value file from excel to be used and */
/* tells what variables to include */

data ASU2;
infile 'C:\Documents and Settings\owner\My Documents\SAS files\
ASU2andRSVandFLU.csv' DSD firstobs=2;
input week $2. ASU2 RSV FLU ;
run;

/* Create a new variable STATEORD that contains the */
/* numerical ordering of observations from the original */
/* data set ONE. */

data add_n;
set ASU2 end=last;
wk_ord=_n_;
run;

/* Create the control data set STNAME using the ADDN */
/* data set. The control data set contains the */
/* following required variables: */
/* START contains the unformatted values */
/* LABEL contains the formatted values (state names) */
/* FMTNAME= contains the name of the format (DATAORD) */
/* TYPE= contains the type of the format, N for numeric */
/* or C for character variables. */
```

```

data wk_name;
  set add_n(rename=(wk_ord=start week=label));
  fmtname='dataord';
  type='N';
  keep fmtname label start type RSV FLU;
run;

proc format cntlin=wk_name;
run;

/* Creates time series plots of ASU2 and RSV patients */

goptions reset=all;

proc gplot data=add_n;
symbol1 i=spline v=circle h=1 c=red;      /* i is to join points */
symbol2 i=spline v=circle h=1 c=blue;     /* v is type of point */
                                           /* h is size of point */
plot ASU2 * wk_ord =1
     RSV * wk_ord =2 /
     overlay
     haxis=axis1
     vaxis=axis2
     legend=legend1;
format wk_ord dataord.;

title 'Asthma Patients Under 2 years and RSV Records by Week';
title2 'Starting week 14 2001 to week 13 2003';

axis1 offset=(2,2) minor=(n=1) order=(1 to 108 by 4) label=('Week');
axis2 label=('Series');

legend1 label=none position=(top left inside);

run;
quit;

CROSS-CORRELATION PLOTS PROGRAM:

/* Cross ACF of difASU2 and difRSV data */

ods trace on; /* Putting ODS trace on will identify the names of */
              /* of all the output given in the procedure */

proc arima data=ASU2andRSVandFLU; /* computes cross-correlation of 2 time series */
identify var=logASU2(1) crosscorr=ONRSVPos(1); /* ASU2 x RSV */
ods output corrgraph=crosscorr_RSV;
/* ODS OUTPUT statement will put the selected outputs into useable */
/* data sets */
ods select corrgraph;

```



```

run;

proc arima data=ASU2andRSVandFLU;
identify var=ONRSVPos(1) crosscorr=logASU2(1); /* RSV x ASU2 */
ods output corrgraph=crosscorr2_RSV;
ods select corrgraph;
ods trace off;
run;

/* Preparing data sets for plotting cross-correlations */

data crosscorr_RSV;
set crosscorr_RSV;
obs=_n_;
run;
data crosscorr2_RSV;
set crosscorr2_RSV;
obs=_n_;
run;
data crosscorr_diflogASU2xdifRSV;
set crosscorr_RSV;
keep lag correlation obs;
if obs GE 50;
run;
data crosscorr_difRSVxdiflogASU2;
set crosscorr2_RSV;
keep lag correlation obs;
if obs GE 50;
run;

/* Plotting Cross ACF of diflogASU2 x difRSV */

goptions reset=all;
proc gplot data=crosscorr_diflogASU2xdifRSV;
symbol i=needle c=blue width=7;
plot correlation * lag /
haxis=axis1
vaxis=axis2
vref= -.19219 .19219 lvref=3;

title 'Cross ACF of differenced logASU2 and RSV data';
title2 'logASU2(1) x RSV(1)';

axis1 label=('Lag') order=(0 to 25 by 1);
axis2 label=('ACF') order=(-1 to 1 by 0.2);

run;

goptions reset=all;
proc gplot data=crosscorr_difRSVxdiflogASU2;
symbol i=needle c=blue width=7;

```

```

plot correlation * lag /
haxis=axis1
vaxis=axis2
vref= -.19219 .19219 lvref=3;

title 'Cross ACF of differenced RSV and logASU2 data';
title2 'RSV(1) x logASU2(1)';

axis1 label=('Lag') order=(0 to 25 by 1);
axis2 label=('ACF') order=(-1 to 1 by 0.2);

run;

PRELIMINARY MODELLING:

/* Using the UCM (Unobserved Component Models) procedure */
/* (Structural Models) for time series data. */
/* Not including a component in the Proc Ucm implies that */
/* that component is not in the model of the time series. */
/* If the component in the model is not stochastic, the */
/* error variance is set to zero and not estimated. */

PROC UCM DATA=ASU2ANDRSVANDFLU;
ID DATE INTERVAL=WEEK; /* ID - specifies date or time variable */
/* INTERVAL - indicates measurement spacing */
MODEL LOGASU2; /* specifies the dependent series that you */
IRREGULAR variance=0 noest; /* specifies the irregular component (epsilon)*/
LEVEL variance=0 noest; /* specifies the level component (mu) */
SLOPE variance=0 noest; /* specifies the slope component (beta) */
CYCLE PRINT=SMOOTHED; /* specifies the cycle component (psi) */
SEASON length=52; /* specifies the season component length */
ESTIMATE ; /* estimates the parameters */

FORECAST OUTFOR=modelFOR lead=0 PRINT=DECOMP;

/* requests series forecasts and forecasts of the sum of components */
/* print=decomp requests the printing of the smoothed trend (mu) and the */
/* trend plus seasonal (mu + gamma) */

ODS output fitstatistics=fitstatistics ;
ODS output fitsummary=fitsummary;
ODS output parameterestimates=parameterestimates;
ods select parameterestimates fitsummary fitstatistics componentsignificance;

title 'UCM of LOGASU2';
RUN;

/* Preparing data sets for Residual Tests */;

```



```

data fitstatistics;
set fitstatistics;
if _n_=1 or _n_=5 or _n_=7;
keep fitstatistic value;
run;

data fitsummary;
set fitsummary;
if _n_=4;
keep fitstatistic value;
run;

data fits;
set fitsummary fitstatistics;
run;
proc transpose data=fits out=fits_transpose;
run;

data out;
set fits_transpose;
label col1='AIC' col2='MSE' col3='R-squared' col4='R.R-squared';
drop _name_;
run;

data model_residuals;
set modelfor;
std_res= residual/std;
std_res_sq=std_res**2;
keep date residual std std_res std_res_sq;
run;

proc timeseries data=model_residuals outcorr=ACF_model_residuals;
var residual;
run;
proc means data=model_residuals n nmiss mean;
var std_res;
ods output summary=statsummary;
ods select summary;
run;

/* Q statistic (Ljung-Box) where the # of d.f. for the Chi-square should be */
/* P-n*, where n*=n-1 where n is the number of hyperparameters */
/* (i.e. the number of parameters estimated by the model). */

proc arima data=model_residuals;
identify var=std_res;
ods output chisqauto=Qtest;
ods select chisquto;
run;

```

```

DATA Qtest;
set Qtest;
DFnew = tolags-(n-1);
aluel = 1-probchi(chisq,dfnew);
keep tolags chisq DFnew pvalue;
run;

DATA Qtest;
set Qtest;
label DFnew='DF'
pvalue='P-value';
run;

/* Heteroscedasticity Test: h=T*/3, T*=T-d, d=no. of non-stationary components */

ods output stat_pvalue=heteroscedasticity;
proc iml;
use statsummary;
read all var {std_res_N std_res_NMiss} into stats[colname=labels];
no_obs=stats[1,1];
d=stats[1,2];
T=no_obs+d;
Tstar=T-d;
m=floor(Tstar/3);
use model_residuals;
read all var {std_res_sq} into res[colname=labels];
num=sum(res[T-m+1:T,]);
den=sum(res[d+1:d+m,]);
stat=num/den;
pvalue=1-probf(stat,m,m);
title 'Heteroscedasticity Test';
print stat pvalue;

/* Normality test of the Residuals */

ods output stat_pvalue=bowmanshenton;

proc iml;
use statsummary;
read all var {std_res_N std_res_NMiss std_res_mean} into stats[colname=labels];
no_obs=stats[1,1];
d=stats[1,2];
T=no_obs+d;
Tstar=T-d;
use model_residuals;
read all var {std_res} into res[colname=labels];
res_mean=res-stats[1,3];
res_mean2=res_mean##2;
res_mean3=res_mean##3;
res_mean4=res_mean##4;
sigma=sqrt((sum(res_mean2[d+1:T,]))/(t-d));

```

```

skew=(1/sigma##3)*((sum(res_mean3[d+1:T,]))/(tstar));
kurt=(1/sigma##4)*((sum(res_mean4[d+1:T,]))/(tstar));
term1=(Tstar/6)*(skew**2);
term2=(Tstar/24)*((kurt-3)**2);
stat=term1 + term2;
pvalue=1-probchi(stat,2);
title 'Bowman-Shenton Test';
print stat pvalue;

data residualtest;
set heteroscedasticity bowmanshenton shapirowilks;
run;

ods output retest=retest;
proc iml;
use residualtest;
read all into retest[colname=labels];
Test={'Heteroscedasticity', 'Bowman-Shenton', 'Shapiro-Wilk'};
Name={'Statistic' 'P-value'};
use Qtest;
read all into qtest[colname=labels];
title 'Residual Tests';
print qtest;
print retest[r=test c=name];
quit;
data retest;
set retest;
label rowname='Test';
run;

/* Defines the template style for which the output will be displayed */

proc template;
    define style styles.output;
        parent=styles.rtf;
    style table from table /
        tagattr='align="left" style="position:relative;top:.2in"';
    style systemtitle from systemtitle /
        protectspecialchars=off;
    style Data from Data /
        font_size=2
        Just=c;
end;
run;

/* Sets the data sets to be outputted in the given template form */
/* This will produce the three charts similar to Tables 4.2,4.11,4.12 */
/* in one output file */

options nodate nonumber;

```

```

ods listing close;
ods rtf file='output.rtf' style=styles.output startpage=yes bodytitle;
title '\b\i0 Fit Statistics';
data _null_;
file print ods;
set out;
put _ods_;
run;
ods rtf startpage=no;
title '\b\i0 Ljung-Box Residual Test';
data _null_;
file print ods;
set Qtest;
put _ods_;
run;
ods rtf startpage=no;
title '\b\i0 More Residual Tests';
data _null_;
file print ods;
set retest;
put _ods_;
run;
ods _all_ close;
ods listing;

/* Plotting residuals of LOGASU2 model */

GOPTIONS RESET=ALL;
proc gplot data=model_residuals;
    symbol1 i=join v=circle h=0 C=BLUE width=1;
    plot RESIDUAL * date /
    overlay
    haxis=axis1
    vaxis=axis2;

title 'Residual Plot of log(ASU2 Model';

axis1 label=('Date')
    order=('1APR01'd to '1APR03'd by year);
axis2 label=('Residuals');

run;

/* Plotting ACF of residuals of LOGASU2 model */

goptions reset=all;
proc gplot data=ACF_model_residuals;
symbol i=needle c=blue width=7;
plot ACF * lag /
overlay
haxis=axis1

```

```

vaxis=axis2;
vref=-.19219 .19219 lvref=3;

title 'Residual ACF of log(ASU2) Model';

axis1 label=('Lag');
axis2 label=('ACF') order=(-1 to 1 by 0.2);

run;

MODELLING RSV AS A PREDICTOR:

/* Proc expand computes transformations of the independent variable */

proc expand data=ASU2andRSVandFLU out=RSV method=none;
    id date;
    convert ONRSVPcs=dif1RSV / transform = (dif 1);
    convert ONRSVPcs=lag1RSV / transform = (lag 1);
    convert ONRSVPcs=lag2RSV / transform = (lag 2);
    convert ONRSVPcs=RSVlead1 / transform =( lead 1 );
    convert ONRSVPcs=RSVlead2 / transform =( lead 2 );
run;

/* Prepares data sets for the predictor modeling by removing missing values */

data RSV1;
set RSV;
keep date logASU2 dif1RSV lag1RSV;
if _n_ =1 then delete;
run;

data RSV2;
set RSV;
lagdifRSV=lag(dif1RSV);
keep date logASU2 lagdifRSV lag2RSV;
if _n_ LE 2 then delete;
run;

data RSV3;
set RSV;
lag2difRSV=lag2(dif1RSV);
keep date logASU2 lag2difRSV;
if _n_ LE 3 then delete;
run;

data RSVlead1;
set RSV;
keep date logASU2 RSVlead1;
if _n_ LT 104;
run;

```

```

data RSVlead2;
set RSV;
keep date logASU2 RSVlead2;
if _n_ LT 103;
run;

data RSV_dif1RSV;
set RSV;
RSV_dif1RSV=ONRSVPos*dif1RSV;
keep date logASU2 ONRSVPos dif1RSV RSV_dif1RSV;
if dif1RSV = . then delete;
run;

/* Program that will run all eight transformations and the final */
/* output will be that similar to Table 4.13 */

PROC UCM Data=RSV;
ID DATE INTERVAL=WEEK;
MODEL LOGASU2=ONRSVPos;
IRREGULAR variance=0 noest;
LEVEL variance=0 noest;
CYCLE PRINT=SMOOTH;
ESTIMATE ;
FORECAST OUTFOR=modelFOR lead=0 PRINT=DECOMP;

ODS output fitstatistics=fitstatistics ;
ODS output fitsummary=fitsummary;
ODS output parameterestimates=parameterestimates;
ods select parameterestimates fitsummary fitstatistics
      componentsignificance;

title 'UCM of ASU2 with RSV as a predictor';
RUN;

data fitstatistics;
set fitstatistics;
if _n_=1 or _n_=5 or _n_=7;
keep fitstatistic value;
run;
data fitsummary;
set fitsummary;
if _n_=4;
keep fitstatistic value;
run;
data fits;
set fitsummary fitstatistics;
run;
proc transpose data=fits out=fits_transpose;
run;

```

```

data estimates;
set parameterestimates;
if _n_=4;
keep component probt estimate ;
run;
data out_RSV1;
merge estimates fits_transpose;
label col1='AIC' col2='MSE' col3='R-squared' col4='R.W. R-squared' probt='p-value';
drop _name_;
run;

```

```

PROC UCM Data=RS1;
ID DATE INTERVAL=WEEK;
MODEL LOGASU2=dif1RSV;
IRREGULAR variance=0 noest;
LEVEL variance=0 noest;
CYCLE PRINT=SMOOTH;
ESTIMATE ;
FORECAST OUTFOR=modelFOR lead=0 PRINT=DECOMP;

```

```

ODS output fitstatistics=fitstatistics ;
ODS output fitsummary=fitsummary;
ODS output parameterestimates=parameterestimates;
ods select parameterestimates fitsummary fitstatistics
      components:gnificance;

```

```

title 'UCM of ASU2 with dif1RSV as a predictor';
RUN;

```

```

data fitstatistics;
set fitstatistics;
if _n_=1 or _n_=5 or _n_=7;
keep fitstatistic value;
run;
data fitsummary;
set fitsummary;
if _n_=4;
keep fitstatistic value;
run;
data fits;
set fitsummary fitstatistics;
run;
proc transpose data=fits out=fits_transpose;
run;
data estimates;
set parameterestimates;
if _n_=4;
keep component probt estimate ;

```

```

run;
data out_RSV2;
merge estimates fits_transpose;
label col1='AIC' col2='MSE' col3='R-squared' col4='R.W. R-squared' probt='p-value';
drop _name_;
run;

PROC UCM Data=RSV;
ID DATE INTERVAL=WEEK;
MODEL LOGASU2=lag1RSV;
IRREGULAR variance=0 noest;
LEVEL variance=0 noest;
CYCLE PRINT=SMOOTH;
ESTIMATE ;
FORECAST OUTFOR=1modelFOR lead=0 PRINT=DECOMP;

ODS output fitstatistics=fitstatistics ;
ODS output fitsummary=fitsummary;
ODS output parameterestimates=parameterestimates;
ods select parameterestimates fitsummary fitstatistics
        componentsignificance;

title 'UCM of ASU2 with lag1RSV as a predictor';
RUN;

data fitstatistics;
set fitstatistics;
if _n_=1 or _n_=5 or _n_=7;
keep fitstatistic value;
run;
data fitsummary;
set fitsummary;
if _n_=4;
keep fitstatistic value;
run;
data fits;
set fitsummary fitstatistics;
run;
proc transpose data=fits out=fits_transpose;
run;
data estimates;
set parameterestimates;
if _n_=4;
keep component probt estimate ;
run;
data out_RSV3;
merge estimates fits_transpose;
label col1='AIC' col2='MSE' col3='R-squared' col4='R.W. R-squared' probt='p-value';
drop _name_;

```



```
run;
```

```
PROC UCM Data=RSV;  
ID DATE INTERVAL=WEEK;  
MODEL LOGASU2=lag2RSV;  
IRREGULAR variance=0 noest;  
LEVEL variance=0 noest;  
CYCLE PRINT=SMOOTH;  
ESTIMATE ;  
FORECAST OUTFOR=modelFOR lead=0 PRINT=DECOMP;
```

```
ODS output fitstatistics=fitstatistics ;  
ODS output fitsummary=fitsummary;  
ODS output parameterestimates=parameterestimates;  
ods select parameterestimates fitsummary fitstatistics  
    componentsignificance;
```

```
title 'UCM of ASU2 with lag2RSV as a predictor';  
RUN;
```

```
data fitstatistics;  
set fitstatistics;  
if _n_=1 or _n_=5 or _n_=7;  
keep fitstatistic value;  
run;  
data fitsummary;  
set fitsummary;  
if _n_=4;  
keep fitstatistic value;  
run;  
data fits;  
set fitsummary fitstatistics;  
run;  
proc transpose data=fits out=fits_transpose;  
run;  
data estimates;  
set parameterestimates;  
if _n_=4;  
keep component probt estimate ;  
run;  
data out_RSV4;  
merge estimates fits_transpose;  
label col1='AIC' col2='MSE' col3='R-squared' col4='R.W. R-squared' probt='p-value';  
drop _name_;  
run;
```

```
PROC UCM Data=RSV;  
ID DATE INTERVAL=WEEK;
```

```

MODEL LOGASU2=lagdifRSV;
IRREGULAR variance=0 noest;
LEVEL variance=0 noest;
CYCLE PRINT=SMOOTH;
ESTIMATE ;
FORECAST OUTFOR=modelFOR lead=0 PRINT=DECOMP;

```

```

ODS output fitstatistics=fitstatistics ;
ODS output fitsummary=fitsummary;
ODS output parameterestimates=parameterestimates;
ods select parameterestimates fitsummary fitstatistics
      componentsignificance;

```

```

title 'UCM of ASU2 with lagdifRSV as a predictor';
RUN;

```

```

data fitstatistics;
set fitstatistics;
if _n_=1 or _n_=5 or _n_=7;
keep fitstatistic value;
run;
data fitsummary;
set fitsummary;
if _n_=4;
keep fitstatistic value;
run;
data fits;
set fitsummary fitstatistics;
run;
proc transpose data=fits out=fits_transpose;
run;
data estimates;
set parameterestimates;
if _n_=4;
keep component probt estimate ;
run;
data out_RSV5;
merge estimates fits_transpose;
label col1='AIC' col2='MSE' col3='R-squared' col4='R.W. R-squared' probt='p-value';
drop _name_;
run;

```

```

PROC UCM Data=RSV;
ID DATE INTERVAL=WEEK;
MODEL LOGASU2=lag2difRSV;
IRREGULAR variance=0 noest;
LEVEL variance=0 noest;
CYCLE PRINT=SMOOTH;

```

```
ESTIMATE ;
FORECAST OUTFOR=modelFOR lead=0 PRINT=DECOMP;
```

```
ODS output fitstatistics=fitstatistics ;
ODS output fitsummary=fitsummary;
ODS output parameterestimates=parameterestimates;
ods select parameterestimates fitsummary fitstatistics
      componentsignificance;
```

```
title 'UCM of ASU2 with lag2difRSV as a predictor';
RUN;
```

```
data fitstatistics;
set fitstatistics;
if _n_=1 or _n_=5 or _n_=7;
keep fitstatistic value;
run;
data fitsummary;
set fitsummary;
if _n_=4;
keep fitstatistic value;
run;
data fits;
set fitsummary fitstatistics;
run;
proc transpose data=fits out=fits_transpose;
run;
data estimates;
set parameterestimates;
if _n_=4;
keep component probt estimate ;
run;
data out_RSV6;
merge estimates fits_transpose;
label col1='AIC' col2='MSE' col3='R-squared' col4='R.W. R-squared' probt='p-value';
drop _name_;
run;
```

```
PROC UCM Data=RSV;
ID DATE INTERVAL=WEEK;
MODEL LOGASU2=RSVlead1;
IRREGULAR variance=0 noest;
LEVEL variance=0 noest;
CYCLE PRINT=SMOOTH;
ESTIMATE ;
FORECAST OUTFOR=modelFOR lead=0 PRINT=DECOMP;
```

```
ODS output fitstatistics=fitstatistics ;
```



```

ODS output fitsummary=fitsummary;
ODS output parameterestimates=parameterestimates;
ods select parameterestimates fitsummary fitstatistics
      componentsignificance;

title 'UCM of ASU2 with RSVlead1 as a predictor';
RUN;

data fitstatistics;
set fitstatistics;
if _n_=1 or _n_=5 or _n_=7;
keep fitstatistic value;
run;
data fitsummary;
set fitsummary;
if _n_=4;
keep fitstatistic value;
run;
data fits;
set fitsummary fitstatistics;
run;
proc transpose data=fits out=fits_transpose;
run;
data estimates;
set parameterestimates;
if _n_=4;
keep component probt estimate ;
run;
data out_RSV7;
merge estimates fits_transpose;
label col1='AIC' col2='MSE' col3='R-squared' col4='R.W. R-squared' probt='p-value';
drop _name_;
run;

```

```

PROC UCM Data=RSV;
ID DATE INTERVAL=WEEK;
MODEL LOGASU2=RSVlead2;
IRREGULAR variance=0 noest;
LEVEL variance=0 noest;
CYCLE PRINT=SMOOTH;
ESTIMATE ;
FORECAST OUTFOR=modelFOR lead=0 PRINT=DECOMP;

```

```

ODS output fitstatistics=fitstatistics ;
ODS output fitsummary=fitsummary;
ODS output parameterestimates=parameterestimates;
ods select parameterestimates fitsummary fitstatistics
      componentsignificance;

```

```

title 'UCM of ASU2 with RSVlead2 as a predictor';
RUN;

data fitstatistics;
set fitstatistics;
if _n_=1 or _n_=5 or _n_=7;
keep fitstatistic value;
run;
data fitsummary;
set fitsummary;
if _n_=4;
keep fitstatistic value;
run;
data fits;
set fitsummary fitstatistics;
run;
proc transpose data=fits out=fits_transpose;
run;
data estimates;
set parameterestimates;
if _n_=4;
keep component probt estimate ;
run;
data out_RSV8;
merge estimates fits_transpose;
label col1='AIC' col2='MSE' col3='R-squared' col4='R.W. R-squared' probt='p-value';
drop _name_;
proc print;
run;

data final_predictors;
set out_RSV1 out_RSV2 out_RSV3 out_RSV4 out_RSV5 out_RSV6 out_RSV7 out_RSV8;
Adj_Pvalue=probt * 8;
run;

proc template;
    define style styles.output;
        parent=styles.rtf;
    style table from table /
        tagattr='align="left" style="position:relative;top:.2in"';
    style systemtitle from systemtitle /
        protectspecialchars=off;
    Style Data from Data /
    font_size=2
    Just=c;

end;
run;

```

```
options nodate nonumber;
ods listing close;
ods rtf file='output.rtf' style=styles.output startpage=yes bodytitle;
title '\b\i0 Fit Statistics';
data _null_;
file print ods;
set final_predictors;
put _ods_;
run;

ods _all_ close;
ods listing;
```

Appendix B: STAMP Output

This appendix includes a sample of the output given by the STAMP program. This output shows only the initial model and the first two transformations of the RSV predictor. The remaining predictors are done in the exact same way.

RTIU2:

MODEL:

logRTIU2 = level + irregular + lag 1 + dummy1 (2001/52) + dummy2 (2002/52) +
dummy3 (2003/1) + dummy4 (2003/2)

*/ Original model */

Method of estimation is Maximum likelihood

The present sample is: 2001 (16) to 2003 (13)

Equation 110.

LRTIU2 = Level + Expl vars + Interv + Irregular

Estimation report

Model with 1 parameters (1 restrictions).

Parameter estimation sample is 2001.16 - 2003.13. (T = 102).

Log-likelihood kernel is 1.876516.

No estimation done.

Eq 110 : Diagnostic summary report.

Estimation sample is 2001.16 - 2003.13. (T = 102, n = 101).

Log-Likelihood is 200.142 (-2 LogL = -400.284).

Prediction error variance is 0.0134698

Summary statistics

| | LRTIU2 |
|----------------|-----------|
| Std.Error | 0.11606 |
| Normality | 1.2022 |
| H(33) | 0.90893 |
| r(1) | 0.28253 |
| r(9) | -0.033005 |
| DW | 1.3351 |
| Q(9, 9) | 10.920 |
| R ² | 0.93877 |

Eq 110 : Estimated variances of disturbances.

Component LRTIU2 (q-ratio)
Irr 0.014284 (1.0000)

Eq 110 : Estimated standard deviations of disturbances.

Component LRTIU2 (q-ratio)
Irr 0.11951 (1.0000)

Eq 110 : Estimated coefficients of final state vector.

| Variable | Coefficient | R.m.s.e. | t-value |
|----------|-------------|----------|------------------|
| Lvl | 0.32720 | 0.18286 | 1.7893 [0.0766] |

Eq 110 : Estimated coefficients of explanatory variables.

| Variable | Coefficient | R.m.s.e. | t-value |
|-------------|-------------|----------|-------------------|
| LRTIU2_1 | 0.95200 | 0.026659 | 35.711 [0.0000] |
| Irr 2001.52 | 0.63087 | 0.12058 | 5.232 [0.0000] |
| Irr 2002.52 | 0.62745 | 0.12108 | 5.1821 [0.0000] |
| Irr 2003. 1 | -0.40421 | 0.12410 | -3.2572 [0.0015] |
| Irr 2003. 2 | -0.41148 | 0.12159 | -3.3841 [0.0010] |

Normality test for Residual LRTIU2

| | |
|--------------------|--------------------|
| Sample Size | 101 |
| Mean | 0.231958 |
| Std.Devn. | 0.946938 |
| Skewness | 0.014369 |
| Excess Kurtosis | 0.244401 |
| Minimum | -2.645569 |
| Maximum | 2.828579 |
| Skewness Chi^2(1) | 0.0034754 [0.9530] |
| Kurtosis Chi^2(1) | 0.25137 [0.6161] |
| Normal-BS Chi^2(2) | 0.25485 [0.8804] |
| Normal-DH Chi^2(2) | 1.2022 [0.5482] |

Goodness-of-fit results for Residual LRTIU2

| | |
|--|-----------|
| Prediction error variance (p.e.v) | 0.013470 |
| Prediction error mean deviation (m.d) | 0.010178 |
| Ratio p.e.v. / m.d in squares | 1.114919 |
| Coefficient of determination R2 | 0.938770 |
| ... based on differences RD2 | 0.468374 |
| ... based on diff around seas mean RS2 | -3.028673 |
| Information criterion of Akaike AIC | -4.189656 |
| ... of Schwartz (Bayes) BIC | -4.035246 |

Chi^2(9) = 10.92 [0.2812]

F(33, 33) = 0.90893 [0.6072]

Normal-BS Chi^2(2) = 0.25485 [0.8804]

*/ RSV PREDICTORS */

1. ONRSVPos

Method of estimation is Maximum likelihood
The present sample is: 2001 (16) to 2003 (13)

Method of estimation is Maximum likelihood
The present sample is: 2001 (16) to 2003 (13)

Equation 113.

LRTIU2 = Level + Expl vars + Interv + Irregular

Estimation report
Model with 1 parameters (1 restrictions).
Parameter estimation sample is 2001.16 - 2003.13. (T = 102).
Log-likelihood kernel is 1.876516.
No estimation done.

Eq 113 : Diagnostic summary report.

Estimation sample is 2001.16 - 2003.13. (T = 102, n = 101).
Log-Likelihood is 192.097 (-2 LogL = -384.195).
Prediction error variance is 0.0132997

Summary statistics

| | LRTIU2 |
|----------------|-----------|
| Std.Error | 0.11532 |
| Normality | 2.6269 |
| H(33) | 0.91830 |
| r(1) | 0.29749 |
| r(9) | -0.042145 |
| DW | 1.3037 |
| Q(9, 9) | 11.953 |
| R ² | 0.93954 |

Eq 113 : Estimated variances of disturbances.

| Component | LRTIU2 (q-ratio) |
|-----------|--------------------|
| Irr | 0.014252 (1.0000) |

Eq 113 : Estimated standard deviations of disturbances.

| Component | LRTIU2 (q-ratio) |
|-----------|-------------------|
| Irr | 0.11938 (1.0000) |

Eq 113 : Estimated coefficients of final state vector.

| Variable | Coefficient | R.m.s.e. | t-value |
|----------|-------------|----------|------------------|
| Lvl | 0.54307 | 0.26780 | 2.0279 [0.0452] |

Eq 113 : Estimated coefficients of explanatory variables.

| Variable | Coefficient | R.m.s.e. | t-value |
|-------------|-------------|------------|-------------------|
| LRTIU2_1 | 0.91842 | 0.040462 | 22.699 [0.0000] |
| ONRSVPos | 0.00031766 | 0.00028817 | 1.1023 [0.2729] |
| Irr 2001.52 | 0.59353 | 0.12511 | 4.744 [0.0000] |
| Irr 2002.52 | 0.64310 | 0.12177 | 5.2811 [0.0000] |
| Irr 2003. 1 | -0.37576 | 0.12662 | -2.9677 [0.0037] |
| Irr 2003. 2 | -0.40807 | 0.12150 | -3.3587 [0.0011] |

Normality test for Residual LRTIU2

| | |
|--------------------------------|------------------|
| Sample Size | 101 |
| Mean | 0.234750 |
| Std.Devn. | 0.941003 |
| Skewness | -0.131424 |
| Excess Kurtosis | 0.488163 |
| Minimum | -2.869261 |
| Maximum | 2.763768 |
| Skewness Chi ² (1) | 0.29075 [0.5897] |
| Kurtosis Chi ² (1) | 1.0029 [0.3166] |
| Normal-BS Chi ² (2) | 1.2936 [0.5237] |
| Normal-DH Chi ² (2) | 2.6269 [0.2689] |

Goodness-of-fit results for Residual LRTIU2

| | |
|--|-----------|
| Prediction error variance (p.e.v) | 0.013300 |
| Prediction error mean deviation (m.d) | 0.010026 |
| Ratio p.e.v. / m.d in squares | 1.120172 |
| Coefficient of determination R ² | 0.939543 |
| ... based on differences RD ² | 0.475088 |
| ... based on diff around seas mean RS ² | -2.977792 |
| Information criterion of Akaike AIC | -4.182758 |
| ... of Schwartz (Bayes) BIC | -4.002613 |

2. dif1RSV

Method of estimation is Maximum likelihood
The present sample is: 2001 (16) to 2003 (13)

Equation 114.

LRTIU2 = Level + Expl vars + Interv + Irregular

Estimation report
Model with 1 parameters (1 restrictions).

Parameter estimation sample is 2001.16 - 2003.13. (T = 102).
 Log-likelihood kernel is 1.876516.
 No estimation done.

Eq 114 : Diagnostic summary report.

Estimation sample is 2001.16 - 2003.13. (T = 102, n = 101).
 Log-Likelihood is 198.67 (-2 LogL = -397.339).
 Prediction error variance is 0.0118242

Summary statistics

| | |
|----------------|----------|
| | LRTIU2 |
| Std.Error | 0.10874 |
| Normality | 3.2987 |
| H(33) | 0.66326 |
| r(1) | 0.22870 |
| r(9) | -0.12852 |
| DW | 1.4427 |
| Q(9, 9) | 10.444 |
| R ² | 0.94625 |

Eq 114 : Estimated variances of disturbances.

| | |
|-----------|--------------------|
| Component | LRTIU2 (q-ratio) |
| Irr | 0.012671 (1.0000) |

Eq 114 : Estimated standard deviations of disturbances.

| | |
|-----------|-------------------|
| Component | LRTIU2 (q-ratio) |
| Irr | 0.11256 (1.0000) |

Eq 114 : Estimated coefficients of final state vector.

| Variable | Coefficient | R.m.s.e. | t-value |
|----------|-------------|----------|------------------|
| Lvl | 0.26126 | 0.17318 | 1.5086 [0.1345] |

Eq 114 : Estimated coefficients of explanatory variables.

| Variable | Coefficient | R.m.s.e. | t-value |
|-------------|-------------|------------|-------------------|
| LRTIU2_1 | 0.96191 | 0.025256 | 38.087 [0.0000] |
| dif1RSV | 0.0026493 | 0.00072859 | 3.6362 [0.0004] |
| Irr 2001.52 | 0.37075 | 0.13422 | 2.7623 [0.0068] |
| Irr 2002.52 | 0.62521 | 0.11404 | 5.4824 [0.0000] |
| Irr 2003. 1 | -0.47862 | 0.11866 | -4.0336 [0.0001] |
| Irr 2003. 2 | -0.49986 | 0.11707 | -4.2697 [0.0000] |

Normality test for Residual LRTIU2

| | |
|-------------|----------|
| Sample Size | 101 |
| Mean | 0.228875 |
| Std.Devn. | 0.942449 |

| | | |
|--------------------------------|-----------|----------|
| Skewness | -0.077980 | |
| Excess Kurtosis | 0.591166 | |
| Minimum | -2.525803 | |
| Maximum | 3.004668 | |
| Skewness Chi ² (1) | 0.10236 | [0.7490] |
| Kurtosis Chi ² (1) | 1.4707 | [0.2252] |
| Normal-BS Chi ² (2) | 1.5731 | [0.4554] |
| Normal-DH Chi ² (2) | 3.2987 | [0.1922] |

Goodness-of-fit results for Residual LRTIU2

| | | |
|---------------------------------------|-----|-----------|
| Prediction error variance (p.e.v) | | 0.011824 |
| Prediction error mean deviation (m.d) | | 0.008850 |
| Ratio p.e.v. / m.d in squares | | 1.136533 |
| Coefficient of determination | R2 | 0.946251 |
| ... based on differences | RD2 | 0.533324 |
| ... based on diff around seas mean | RS2 | -2.536477 |
| Information criterion of Akaike | AIC | -4.300354 |
| ... of Schwartz (Bayes) | BIC | -4.120209 |

Chi²(9) = 10.444 [0.3158]

F(33, 33) = 0.66326 [0.8783]

Normal-BS Chi²(2) = 1.5731 [0.4554]

Bibliography

- [1] Brockwell, P. & Davis, R.A. (2002) *Introduction to Time Series and Forecasting* 2nd ed., Springer-Verlag New York Inc., New York.
- [2] Harvey, A.C. (1989) *Forecasting, structural time series models and the Kalman filter*, Cambridge University Press, United Kingdom.
- [3] Koopman, S., Harvey, A.C., Doornik, J.A., & Shephard N. (2000) *Structural Time Series Analyser, Modeller and Predictor* 2nd ed., Timberlake Consultants Press, United Kingdom.
- [4] Scuffham P.A. (2003) Estimating Influenza-Related Hospital Admissions in Children and Adults, *Dis Manage Helath Outcomes*, **11**(4), 259-269.
- [5] Scuffham P.A. (2004) Estimating influenza-Related Hospital Admissions in Older People from GP consultation data, *Vaccine*, **22**, 2853-2862.
- [6] SAS 9.1 *The UCM Procedure*, World Wide Web,
[www.http://support.sas.com/rnd/app/papers/ucm.pdf](http://support.sas.com/rnd/app/papers/ucm.pdf).