Application and Comparison of Time Series Models to AIDS Data

by

Tadesse Haileyesus

(Under the direction of Lynne Billard)

Abstract

A class of space time bilinear (STBL) models have been proposed by Dai and Billard (1998). In this paper we have applied these newly developed STBL models, existing space time autoregressive and moving average (STARMA) models, and the standard nonspatial linear autoregressive and moving average (ARMA) models to a set of reported AIDS data in U.S. and compared the results. The STBL models were found to be the best fit to our AIDS data; a discussion of the findings is given.

Index words: Space Time Bilinear Model, Bilinear Time Series, Maximum Likelihood Estimation, Forecasting, Weighting with Probability Proportional to Size, STBL, STARMA, ARMA, HIV/AIDS

Application and Comparison of Time Series Models to AIDS Data

by

Tadesse Haileyesus

A Thesis Submitted to the Graduate Faculty
of The University of Georgia in Partial Fulfillment
of the
Requirements for the Degree

Master of Science

Athens, Georgia

2002

Application and Comparison of Time Series Models to AIDS Data

by

Tadesse Haileyesus

Approved:

Major Professor: Lynne Billard

Committee: T. N. Sriram
Nancy Lyons

Electronic Version Approved:

Gordhan L. Patel
Dean of the Graduate School
The University of Georgia
May 2002

To my parents, Tsegemariam and Haileyesus, to my beloved wife, Saba, and to my children, Adam, Omega and Nathan.

Acknowledgements

I am grateful to my major professor Dr. Lynne Billard, for her patience, enthusiasm, and encouragement in extended guiding in preparing this paper. It is a privilege to work with her. I wish to thank my advisory committee members, Dr. Nancy Lyons and Dr. T. N. Sriram for their carefully reading and for graciously being in my committee. I also want to thank Mrs. Molly Rema for giving me a peace of mind by her fast and accurate typing, and Dr. Jaxk Reeves for his advise since the beginning of my study.

My main thanks is extended to Faculty and Staff and students of the Department of Statistics, who helped me in many ways. Support from my employer, The University System of Georgia is gratefully acknowledged.

Finally, I want to especially thank my God's gift, my family, parents and friends for their support, without which this paper and my study would not have been possible. Thank you God!

Table of Contents

Chapter 1

Introduction

1.1 Objectives of the Study

The twin objectives of this study are (1) to apply the class of space time bilinear (STBL) models proposed by Dai and Billard (1998) to a set of AIDS data and (2) to compare the results with existing space time autoregressive and moving average (STARMA) models and the standard nonspatial linear autoregressive and moving average (ARMA) models. The main objectives of time series modeling and analyses are to seek an understanding of the dynamic or time-dependent structure of the observations of a single series-*univariate* time series analysis, and thence to ascertain the leading, lagging, and feedback relationships that might exist among several series-*multivariate* time series analyses. Knowledge of the dynamic structure will help produce better forecasts of future observations and help design more optimal control schemes.

As Dai and Billard (1998) discussed, the STBL model is a special form of a multiple bilinear time series which exhibits bilinear behavior on a spatial neighborhood structure. Spatial data are data collected at a number of geographically separate locations. States and regions in a county are very common spatial neighborhoods in epidemic monitoring including in AIDS surveillance. Since spatial time series models deal with processes that exhibit both spatial and temporal dependencies, the data set consisting of AIDS cases in U.S. reported from all regions over time is analyzed and so addresses our main objectives.

This paper is organized as follows. Chapter 2 presents a literature review of ARMA models, STARMA models, and STBL models. Background, the Data Sets, and Autocorrelation and Partial Autocorrelation Functions are discussed in Chapter 3. Chapter 4 deals with the analysis of the data using the three models (of Chapter 2). The major topics cover data sources, missing observations, weighting with probability proportional to size, neighborhood structure, maximum likelihood estimation of the coefficients and variance of error in the models, and forecasting. In Chapter 5, comparisons of the results and conclusions of this paper are given.

This paper is formatted with LaTeX using The University of Georgia thesis style file and the electronic version may be accessible through the following website: http://graduate.gradsch.uga.edu/etd2/library.

1.2 Expected Outcomes of the Research

Dai and Billard (1998) developed a space time model, the STBL model, which can be employed for real data that exhibits both spatial and sudden short spurts of larger than usual values (i.e., shocks features to the underlying time series process). An application to the epidemic surveillance data of the infections disease mumps in U.S. is presented in Billard and Dai (2000) and of wind speeds in Ireland in Dai and Billard (2002). Based on their discussion and findings, it is reasonable to expect (1) the feasibility of applying the STBL model to the modeling of the infections disease AIDS in U.S., and (2) the fits from the STBL model to be preferable than those obtained from existing models.

Chapter 2

Literature Review

2.1 Autoregressive Moving-Average Model

As Box and Jenkins (1976) discussed, the use of time series and dynamic models in three important areas of applications are (1) the *forecasting* of future values, (2) the determination of the *transfer function* of a system, and (3) the design of simple *feed forward and feed back control schemes*. The autoregressive moving-average (ARMA) model or more generally the autoregressive integrated moving-average (ARIMA) model, also often referred to as the Box-Jenkins model procedure, analyzes and forecasts equally spaced univariate time series data, transfer function data and intervention data. The ARIMA model predicts a value in a response time series as a linear combination of its own past values and of past errors also called shocks, innovations or white noises.

The general transfer function model employed by the ARIMA procedure was discussed Box and Jenkins (1976). Aroian (1980), Voss et al. (1980), Oprian et al. (1980), Taneja et al. (1980) and others, have extended the Box-Jenkins results. The standard linear autoregressive moving average (ARMA) model is given by

$$\begin{aligned} Z(t) \;&=\; \sum_{i=1}^{p} \phi_i \; Z(t-i) \\ &+ \; \sum_{j=1}^{q} \; \theta_j e(t-j) \; - \; e(t) \end{aligned} \tag{2.1}$$

where $Z(t)$ is a sequence of observations,

ϕ_i is the autoregressive parameter,

θ_j is the moving average parameter, and

$e(t)$ are assumed to be independently and identically distributed random variables from a normal distribution $N(0, \sigma^2)$.

In the papers of Aroian et al., conditions are discussed for allowable parameters values in such ARMA models to ensure stationarity and invariability for low order cases. We denote ARMA(p, q) with p and q representing the autoregressive order and moving average order of the model, respectively.

Henceforth, in this work, it will be assumed that suitable differencing of the data has been carried out to ensure stationarity pertains, and that suitable transformations have been executed to ensure stability of the variance.

2.2 Space Time Autoregressive and Moving Average Model

Linear autoregressive moving average space time models known as STARMA models have been developed since 1975 (Cliff et al., 1975; Pfeifer and Deutsch, 1980a, 1980b). The STARMA model, denoted by STARMA $(p_{\boldsymbol{\lambda}}, q_{\boldsymbol{\eta}})$, is defined as

$$\begin{aligned} \mathbf{Z}(t) &= \sum_{i=1}^{p} \sum_{m=0}^{\lambda_i} \phi_m^i \mathbf{W}^{(m)} \mathbf{Z}(t-i) \\ &+ \sum_{j=1}^{q} \sum_{n=0}^{\eta_j} \theta_n^j \, \mathbf{W}^{(n)} \mathbf{e}(t-j) + \mathbf{e}(t) \end{aligned} \tag{2.2}$$

where $\mathbf{Z}(t) = [Z_1(t), ..., Z_n(t)]^T$ is an $n \times 1$ stochastic vector process,

$\mathbf{W}^{(m)} = (\mathbf{w}_{kn}^{(m)})$ is the $n \times n$ weighting matrix for spatial order m,

ϕ_m^i is the autoregressive parameter at temporal lag i and spatial lag m,

θ_n^j is the moving average parameter at temporal lag j and spatial lag n,

$\boldsymbol{\lambda}_i = [\lambda_1, ..., \lambda_p]$ where λ_i is the spatial order associated with the ith autoregressive

parameter,

$\boldsymbol{\eta}_j = [\eta_1, ..., \eta_q]$ where η_j is the spatial order associated with the jth moving average parameter,

and $\mathbf{e}(t) = [e_1(t), ..., e_n(t)]^T$ is a sequence of independent and identically distributed vector random variables.

Like the ARMA models, the linear space time models have already appeared in many important areas of applications. Note that when $n = 1$, $W^{(m)} = 0,\ m > 0,\ W^{(0)} = I$, this model reduces to the standard ARMA model as a special case.

2.3 Space Time Bilinear Model

To model nonlinear phenomena such as spatial and temporal processes of earthquakes, acute infectious disease outbreaks, etc., nonlinear models are considered. One such powerful nonlinear time series model is the bilinear (BL) model which emerged as a direct extension of ARIMA models by including cross products of $Z(t)$ and $e(t)$; this was first developed in the context of control theory by Mohler (1973). The analysis of BL models has been considered by Granger and Andersen (1978), Subba Rao (1981), Kim and Billard (1990), and Subba Rao and Gabr (1984), among others. However, this BL model does not allow for spatial dependencies.

Recently a class of space time bilinear (STBL) models designed to model spatial time series data which exhibit bilinear behavior has been proposed by Dai and Billard (1998). Analogous to the multivariate bilinear time series model, the STBL model contains the most general STARMA models (2.2) as one special case when the nonlinear part, i.e., the pure bilinear term is not present. The STBL model also contains the general univariate bilinear BL model as another special case when each

site is not spatially dependent and can be handled separately by letting $n = 1$. Dai and Billard (1998) defined the STBL model as follows:

$$\mathbf{Z}(t) = \sum_{i=1}^{p} \sum_{m=0}^{\lambda_i} \phi_m^i \mathbf{W}^{(m)} \mathbf{Z}(t-i) + \sum_{j=1}^{q} \sum_{n=0}^{\eta_j} \theta_n^j \mathbf{W}^{(n)} \mathbf{e}(t-j)$$

$$+ \sum_{i=1}^{r} \sum_{j=1}^{s} \sum_{m=0}^{\xi_i} \sum_{n=0}^{\mu_j} \beta_{mm}^{ij} [\mathbf{W}^{(m)} \mathbf{Z}(t-i)] \# [\mathbf{W}^{(n)} \mathbf{e}(t-j)] + \mathbf{e}(t) \tag{2.3}$$

where $\mathbf{Z}(t) = [Z_1(t), ..., Z_g(t)]^T$, $t = 0, 1, ...$ is an $g \times 1$ stochastic vector process,

p is the autoregressive order,

q is the moving average order,

r is the autoregressive order in the bilinear term,

s is the moving average order in the bilinear term,

λ_i is the spatial order of the autoregressive term at temporal lag i,

η_j is the spatial order of the moving average term at temporal lag j,

ξ_i is the spatial order of the autoregressive term in the bilinear term at temporal lag i,

μ_j is the spatial order of the moving average term in the bilinear term at temporal lag j,

ϕ_m^i is the autoregressive parameter at temporal lag i and spatial lag m,

θ_n^j is the moving average parameter at temporal lag j and spatial lag n,

β_{mm}^{ij} is the bilinear parameter at temporal lag i and j for the autoregressive and the moving average terms, respectively, and at spatial lag m and n for the autoregressive and moving average terms, respectively,

$\mathbf{W}^{(m)} = (w_{ku}^{(m)})$ is the $g \times g$ weighting matrix at spatial order m, and

$\mathbf{e}(t) = [e_1(t), ..., e_g(t)]^T$ is a sequence of independent and identically distributed vector random variables with

$$E[\mathbf{e}(t)] = 0,$$

$$E[\mathbf{e}(t)\mathbf{e}(t+j)^T] = \begin{cases} G, \ j = 0, \\ 0, j \neq 0, \end{cases}$$

$$E[\mathbf{Z}(t)\mathbf{e}(t+j)^T] = 0, \ j > 0.$$

We write $\mathbf{A}\#\mathbf{B} = (c_{ij})$, where $c_{ij} = a_{ij}b_{ij}$ is defined as matrix element wise multiplication for any matrices $\mathbf{A} = (a_{ij})$ and $\mathbf{B} = (b_{ij})$ of the same size. Analogous to the STARMA model in (2.2), the model of (2.3) is denoted by STBL $(p_{\boldsymbol{\lambda}}, q_{\boldsymbol{\eta}}, r_{\boldsymbol{\xi}}, s_{\boldsymbol{\mu}})$ model of temporal order p, q, r, and s; and spatial order $\boldsymbol{\lambda} = [\lambda_1, ..., \lambda_p]$, $\boldsymbol{\eta} = [\eta_1, ..., \eta_q]$, $\boldsymbol{\xi} = [\xi_1, ..., \xi_r]$, and $\boldsymbol{\mu} = [\mu_1, ..., \mu_p]$.

As Dai and Billard explained, the spatial locations can be regularly or irregularly spaced. An example of regularly spaced spatial locations would be that where observations are placed on a grid or on quadrants of Euclidean space; whereas common examples of irregularly spaced locations would be states or regions in a country. More theoretical results accompanied by examples and numerical illustrations can be found in Dai and Billard (1998, 2002), and Billard and Dai (2000).

Chapter 3

Background and the Data Set

3.1 The Data Set

The data set for this research is the set consisting of the number of acquired immunodeficiency syndrome (AIDS) cases reported in the United States. The AIDS Public Information Data Set is created each year by the Division of HIV/AIDS Prevention, National Center for HIV, STD and TB Prevention, Center for Disease Control and Prevention (CDC) and contains information extracted from CDC's national AIDS surveillance data base. One of the two formats of the AIDS Public Information Data Set consists of a rectangular data file of 16 variables extracted from CDC's national AIDS data base (available at: http://www.cdc.gov/hiv/software/apids.htm). This rectangular data file contains one line of data for each AIDS case reported to CDC. Each line contains a total of 35 columns (for the 16 variables).

The data set for this paper contains the year and month in which CDC received the case report, and region (non-metropolitan area) of residence at diagnosis of AIDS identified since January 1984 through December 1999 (see Appendix B.1). According to the U.S. Bureau of Census, CDC classified the 51 states into four regions (Northeast, Midwest, South and West); see Figure 3.1 and Table 3.1.

Only three reports were missing out of 768 (16 years x 12 months x 4 regions) expected monthly reports amounting to 0.004%. Two of these missing reports (one case each from the Midwest and West in 1984) were surrounded by low counts of

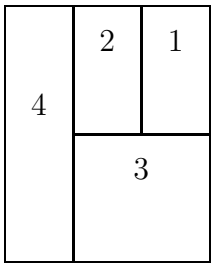
cases (less than 10) and treated as zeros, while the other one (from the West in 1992) was surrounded by high counts and so was handled by taking the average of the four surrounding counts plus a random number between 0 and 9 (to maintain its randomness).

Finally, the data were examined for seasonality and nonstationarity through the autocorrelation and partial autocorrelation functions, and appropriate differencing and transformation methods were employed to convert these data into data that were nonseasonal and stationary; see Section 3.3.

3.2 Weighting with Probability Proportional to Size

The Dai and Billard model of (2.3) requires that the weight matrices $\mathbf{W}^{(m)}$ be known. There are many possible formats for $\mathbf{W}^{(m)}$. In this paper, employing the advantages of sampling with probability proportional to size (PPS), we have developed a probability proportional to size weighting matrix scheme as follows.

Intuitively, it seems more reasonable to put more weight on the neighbor that has more cases than the one with fewer counts (since believed it will exerts more influence). Let us define the neighbor to be the region that shares the common borders. The neighborhood structure can be simplified as follows (see Figure 3.1; the following diagram is not drawn to scale):



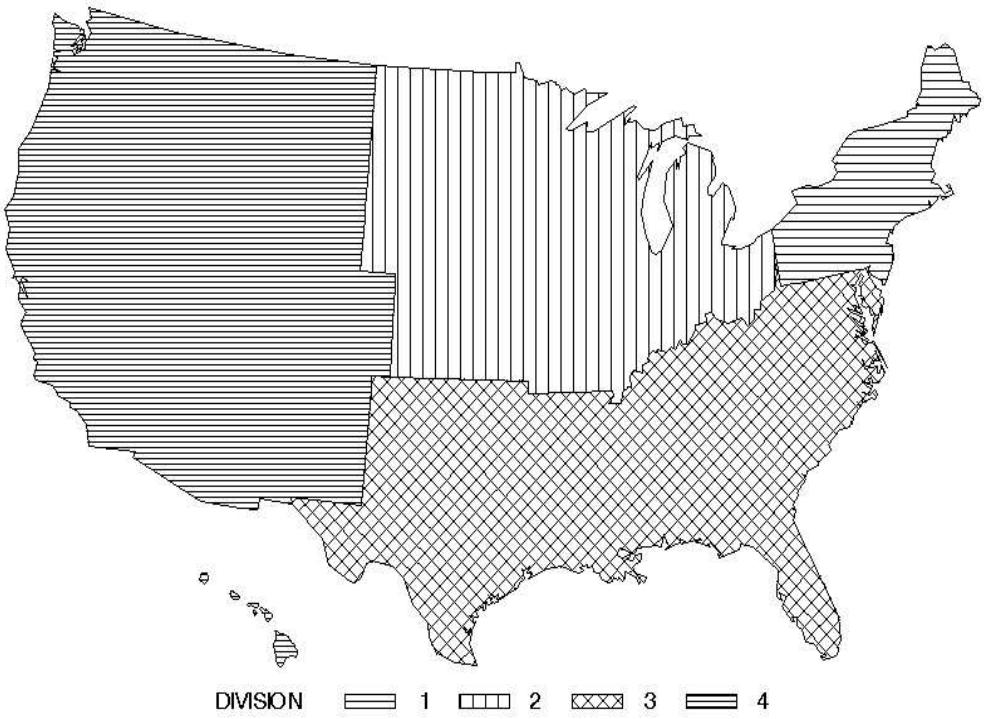


Figure 3.1. U.S. Census Bureau Regions Division Map

Region 1 is the Northeast with 21,834 reported AIDS cases,

Region 2 is the Midwest with 15,349 reported AIDS cases,

Region 3 is the South with 62,104 reported AIDS cases, and

Region 4 is the West with 16,009 reported AIDS cases .

Hence, the first spatial order weighting matrix for the first order neighborhood (see Table 3.2) which identifies the neighbors for each of the regions can be constructed as follows

$$\boldsymbol{W}^{(1)} = \begin{bmatrix} 0.0 & 0.2 & 0.8 & 0.0 \\ 0.2 & 0.0 & 0.6 & 0.2 \\ 0.4 & 0.3 & 0.0 & 0.3 \\ 0.0 & 0.2 & 0.8 & 0.0 \end{bmatrix}.$$

From the first row of the weighting matrix, we can see that weighting with PPS assigns more weight to the South Region (0.8) than to the Midwest Region (0.2) for the same first order neighbors of the Northeast Region 1. This means in particular that Region 1 is likely to be influenced more by Region 3 than it will be by Region 2.

3.3 Autocorrelation and Partial Autocorrelation Functions

The original time series data (shown in Figures 3.2.a, 3.2.b, 3.2.c, and 3.2.d with the solid line for the Northeast Region, Midwest Region, South Region, and West Region, respectively) were transformed to achieve stability in the variance of the underlying model. Each of the well known square root and log transformations was used in separate analyses, one for each region. Since these analyses showed the log transformation to be superior, we report henceforth on that one only.

It is well known that variance is a measure of dispersion and that comparability is achieved if we standardize the autocovariances by dividing them all by covariance's, that is, by transforming them to correlations, which for time series data are referred to as autocorrelations.

The set of autocorrelations is often referred to collectively as the *autocorrelation functions*. A graph of the autocorrelation function, called the correlogram, serves much the same function in time series analysis as does the histogram in sampling.

The approximation for a standard error for the estimated partial autocorrelation function at lag k is based on a null hypothesis that a pure autoregressive Gaussian process of order $k-1$ generated the time series. This standard error is used to produce the approximate 95% confidence intervals depicted by the dots in the plot.

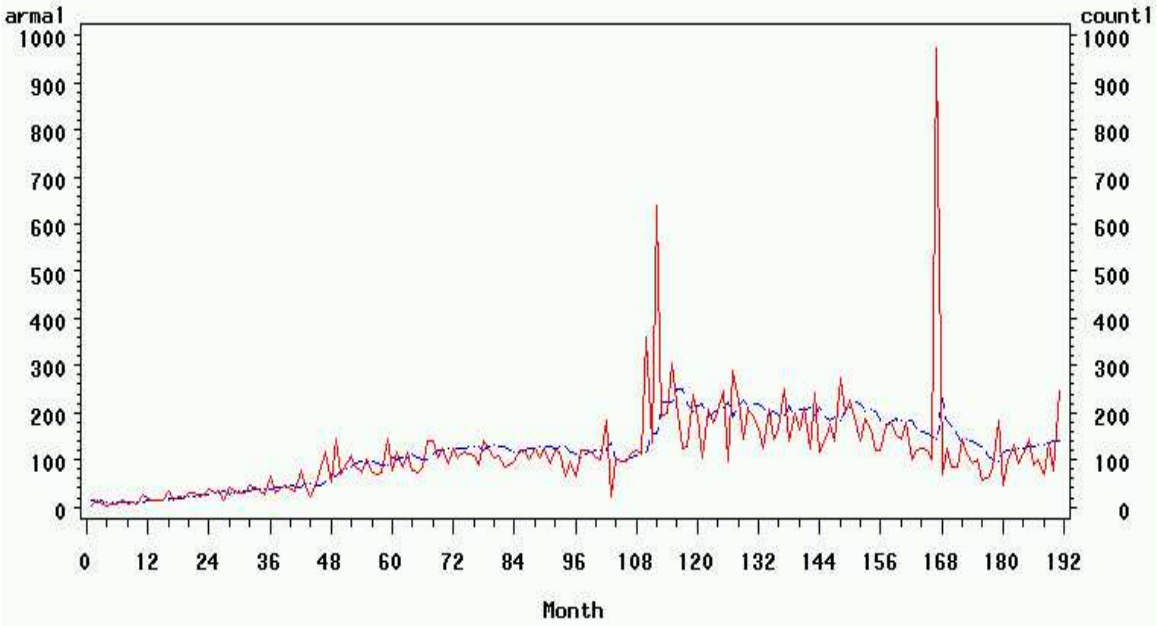


Figure 3.2.a Reported and Predicted AIDS Cases by Month, Northeast Region
Reported - Predicted ... using ARMA

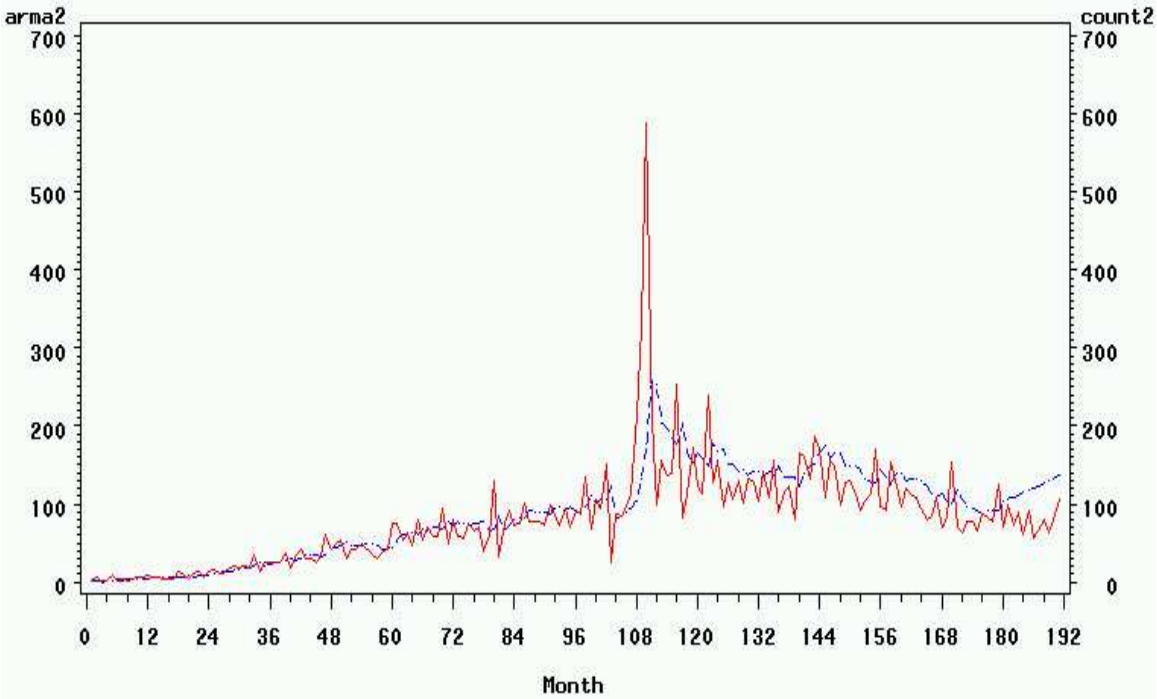


Figure 3.2.b Reported and Predicted AIDS Cases by Month, Midwest Region
Reported - Predicted ... using ARMA

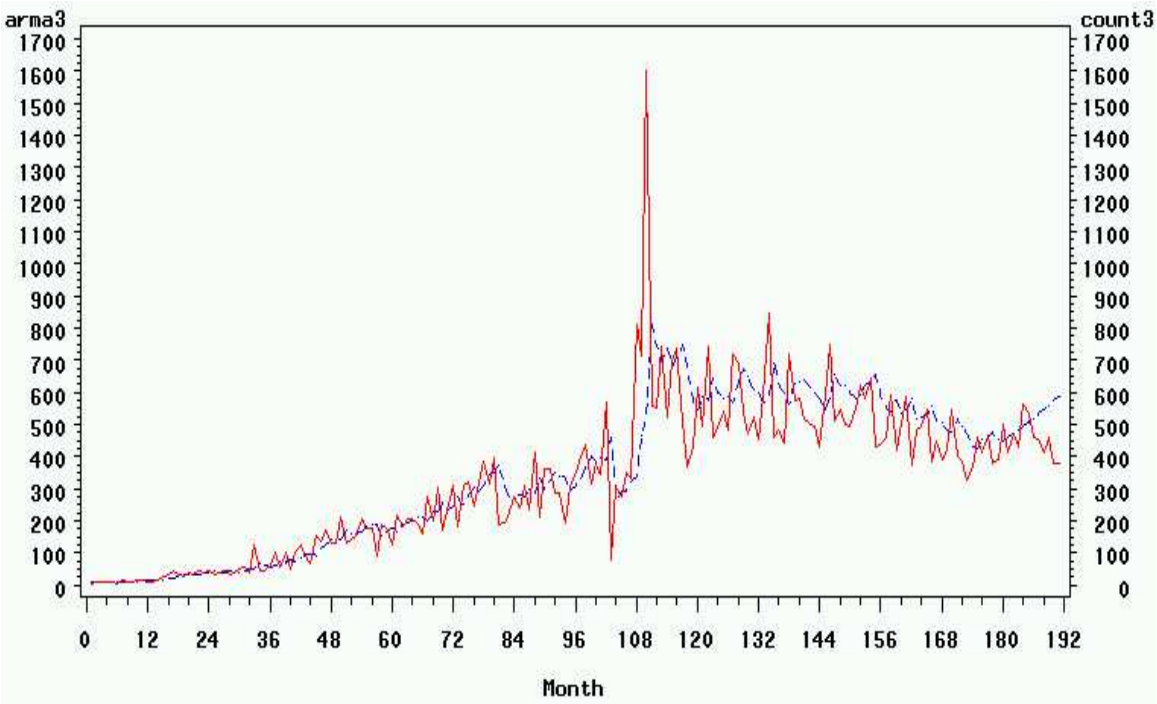


Figure 3.2.c Reported and Predicted AIDS Cases by Month, South Region
Reported - Predicted ... using ARMA

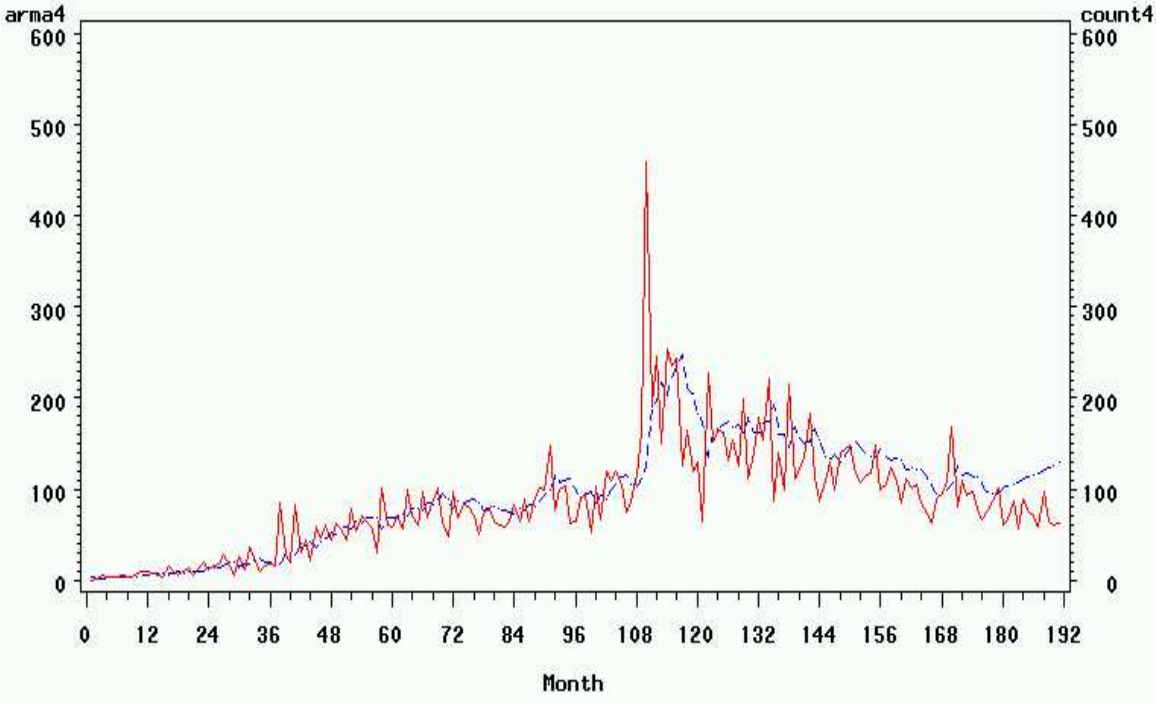


Figure 3.2.d Reported and Predicted AIDS Cases by Month, East Region
Reported - Predicted ... using ARMA

The autocorrelation functions based on the original (log transformed) data (see Figure 3.3.a) decays but not fast enough as we go from lag 0 to lag 24. The partial autocorrelation function is shown in Figure 3.3.b.

Obviously, this suggests the need to apply a differencing operation. Among several potential differencing candidates were considered, the first differencing operation emerged as the most appropriate for our data (followed by the well known lag twelve differencing operation).

Then we have employed this first differencing operation to impose stationarity on the resulting underlying ARIMA procedure which is further discussed in the following chapter.

The ARIMA Procedure

Name of Variable = logct

Mean of Working Series	4.424845
Standard Deviation	0.895545
Number of Observations	192

Autocorrelations

Lag	Covariance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
0	0.802001	1.00000	|**********|
1	0.631781	0.78776	|********** . |
2	0.645688	0.80510	|********** . |
3	0.626543	0.78123	|********** . |
4	0.609946	0.76053	|********** . |
5	0.580270	0.72353	|********** . |
6	0.557290	0.69487	|********** . |
7	0.535624	0.66786	|********** . |
8	0.523781	0.65309	|********** . |
9	0.499600	0.62294	|********** . |
10	0.491294	0.61259	|********** . |
11	0.447346	0.55779	|********** . |
12	0.473493	0.59039	|********** . |
13	0.445094	0.55498	|********** . |
14	0.432446	0.53921	|********** . |
15	0.412203	0.51397	|**********. |
16	0.392369	0.48924	|**********. |
17	0.400161	0.49895	|**********. |
18	0.389311	0.48542	|**********. |
19	0.361446	0.45068	|********* . |
20	0.346335	0.43184	|********* . |
21	0.349497	0.43578	|********* . |
22	0.333840	0.41626	|******** . |
23	0.318184	0.39674	|******** . |
24	0.318184	0.39674	|******** . |

"." marks two standard errors

Figure 3.3.a Autocorrelation Functions, Northeast Log Transformed Data

Partial Autocorrelations

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	0.78776	|. |********** |
2	0.48634	|. |********** |
3	0.24970	|. |***** |
4	0.11921	|. |**. |
5	-0.00784	|. | . |
6	-0.04007	|. *| . |
7	-0.02827	|. *| . |
8	0.03114	|. |* . |
9	0.00354	|. | . |
10	0.04182	|. |* . |
11	-0.10473	|.**| . |
12	0.13210	|. |*** |
13	0.03917	|. |* . |
14	-0.00652	|. | . |
15	-0.04764	|. *| . |
16	-0.07882	|.**| . |
17	0.07485	|. |* . |
18	0.07249	|. |* . |
19	-0.04374	|. *| . |
20	-0.08140	|.**| . |
21	0.04072	|. |* . |
22	-0.00104	|. | . |
23	0.02061	|. | . |
24	0.03774	|. |* . |

Autocorrelation Check for White Noise

To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	683.94	6	<.0001	0.788	0.805	0.781	0.761	0.724	0.695
12	1152.03	12	<.0001	0.668	0.653	0.623	0.613	0.558	0.590
18	1486.64	18	<.0001	0.555	0.539	0.514	0.489	0.499	0.485
24	1719.67	24	<.0001	0.451	0.432	0.436	0.416	0.397	0.397

Figure 3.3.b Partial Autocorrelation Functions, Northeast Log Transformed.

Chapter 4

Application of Time Series Models

4.1 ARMA Models

We have examined the autocorrelation and partial autocorrelation functions of each region by extending the lag up to 48 (the SAS default lag is 24) to obtain a better vision of the underlying ARMA process. As a result the of log transformation and first differencing operation, we can see from our correlogram in Figure 4.1.a that now, the autocorrelations decay quickly as intended (compare Figure 3.3.a and Figure 4.1.a for Northeast region).

Moreover, we can see that the lag $k = 1$ autocorrelation function values are large, and that for lags $k > 1$ the autocorrelation function values are very small, while the partial autocorrelation function values decayed exponentially to zero in Figure 3.3.b and Figure 4.1.b. This suggests that the first order moving average of lag 1 be considered for the underlying model. This pattern prevailed for all four regions. That is, for each region, the tentatively identified model is an ARMA $(0, 1)$ model (or equivalently for the nondifferenced data, an ARIMA $(0, 1, 1)$ model).

The ARIMA Procedure

Period(s) of Differencing	1
Mean of Working Series	0.014645
Standard Deviation	0.566095
Number of Observations	191
Observation(s) eliminated by differencing	1

Autocorrelations

Lag	Covariance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
0	0.320463	1.00000	|********************|
1	-0.190847	-.59553	************| . |
2	0.041931	0.13084	|***. |
3	-0.0037162	-.01160	| . |
4	0.0056567	0.01765	| . |
5	-0.0019888	-.00621	| . |
6	-0.0053784	-.01678	| . |
7	-0.0012196	-.00381	| . |
8	0.0069668	0.02174	| . |
9	-0.016211	-.05059	. *| . |
10	0.035259	0.11002	|** . |
11	-0.055719	-.17387	.***| . |
12	0.044193	0.13790	|***. |
13	-0.012846	-.04009	. *| . |
14	0.0096701	0.03018	|* . |
15	-0.0007864	-.00245	| . |
16	-0.024448	-.07629	. **| . |
17	0.011782	0.03677	|* . |
18	0.020157	0.06290	|* . |
19	-0.015472	-.04828	. *| . |
20	-0.012024	-.03752	. *| . |
21	0.018078	0.05641	|* . |
22	-0.0042889	-.01338	| . |
23	-0.010948	-.03416	. *| . |
24	0.010850	0.03386	|* . |

"." marks two standard errors

Figure 4.1.a Autocorrelation Functions, Northeast Log Transformed and First Differenced Data

Partial Autocorrelations

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	-0.59553	| ********** | . |
2	-0.34682	| ******* | . |
3	-0.19938	| **** | . |
4	-0.08234	| .** | . |
5	-0.01931	| . | . |
6	-0.02639	| . * | . |
7	-0.05274	| . * | . |
8	-0.02453	| . | . |
9	-0.08084	| .** | . |
10	0.07119	| . |* . |
11	-0.10282	| .** | . |
12	-0.05040	| . * | . |
13	0.00012	| . | . |
14	0.06818	| . |* . |
15	0.11044	| . |** . |
16	-0.03830	| . * | . |
17	-0.11538	| .** | . |
18	0.00885	| . | . |
19	0.06748	| . |* . |
20	-0.03398	| . * | . |
21	0.00724	| . | . |
22	-0.00925	| . | . |
23	-0.04470	| . * | . |
24	-0.01429	| . | . |

Autocorrelation Check for White Noise

To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	72.30	6	<.0001	-0.596	0.131	-0.012	0.018	-0.006	-0.017
12	85.49	12	<.0001	-0.004	0.022	-0.051	0.110	-0.174	0.138
18	88.37	18	<.0001	-0.040	0.030	-0.002	-0.076	0.037	0.063
24	90.41	24	<.0001	-0.048	-0.038	0.056	-0.013	-0.034	0.034

Figure 4.1.b Partial Autocorrelation Functions, Northeast Log Transformed and First Differenced Data

After tentative values for the model orders are identified as discussed above, maximum likelihood estimation of the parameter values was carried out by fitting the identified model and then subsequently by over-fitting (by adding one parameter at a time to) the model.

We have considered a total of 24 different models, 8 each for the square root transformed, log transformed and the original (nontransformed) data.

Then, best fits were obtained by taking the AIC criterion (Akaike, 1977) into consideration. In all four regions, the ARIMA (p, d, q) = (0, 1, 1) model as originally identified, had the best fit; i.e., first order moving average model with no autoregressive component. The AIC values were 213.63, 174.08, 93.94 and 171.94, respectively, and the $\hat{\sigma}^2$ were 0.17, 0.14, 0.09 and 0.14 for Northeast, Midwest, South and East Regions, respectively; see Table 4.1 for the summary of the ARIMA estimation procedure.

Hence, the linear autoregressive moving average model Eq. (2.1) for region k is given by

$$z(t) \;=\; \theta_k \; e(t-1) \;\; - e(t) \tag{4.1}$$

where $\hat{\theta}_1 = 0.772$ for the Northeast region, $\hat{\theta}_2 = 0.721$ for the Midwest region, $\hat{\theta}_3 = 0.673$ for the South region, and $\hat{\theta}_4 = 0.722$ for the East region.

Finally, using these best models we have computed the predicted AIDS cases for each region and reverse transformed. The plot of both the reported and predicted values given in Figures 3.2.a, 3.2.b, 3.2.c, and 3.2.d for the Northeast, Midwest, South and East regions, respectively. Further comparative discussion of these fits is deferred to Chapter 5.

4.2 STARMA Models

Following the procedure discussed by Dai and Billard (1998), we have identified the orders of the model parameters for $(p, q, \boldsymbol{\lambda}, \boldsymbol{\eta})$. The autocorrelation function is given by

$$\rho_h(j) = \tau_{ho}(j)/\sqrt{\tau_{hh}(0)\tau_{oo}(0)} \tag{4.2}$$

where $\rho_{hk}(j)$ is the space-time autocovariance function between the hth and kth order neighbors and j time lags apart and is given by

$$\rho_{hk}(j) \;=\; (1/g)Cov(\mathbf{W}^{(h)}\mathbf{z}(t), \mathbf{W}^{(k)}\mathbf{z}(t-j)) = (1/g)tr([\mathbf{W}^{(\mathbf{k})}]^{\mathbf{T}}\mathbf{W}^{(\mathbf{h})}\boldsymbol{\Gamma}(\mathbf{j}))$$

with

$$\boldsymbol{\Gamma}(j) = Cov(\boldsymbol{z}(t), \boldsymbol{z}(t-j))$$

and tr $(\mathbf{A})$ is the trace of the matrix $\mathbf{A}$. The space-time partial autocorrelation function is the coefficient $\phi_{kl}^{'}$ obtained from solving the system of equations

$$\tau_{ho}(j) = \sum_{i=1}^{k}\sum_{l=1}^{\lambda}\phi_{kl}^{'}\tau_{hl}(j-i) \tag{4.3}$$

as l=0, 1, ..., λ for k=1,2, ..., in turn.

The autocorrelation and partial autocorrelation functions given in equation (4.2) and (4.3) are calculated on the log transformed observations are shown in Table 4.2 and Table 4.3, respectively. From these tables, we see that the autocorrelations cutoff at the S=1 time lag while the partial autocorrelations decay exponentially as the lag S increases. Thus, we have identified the time orders to be p=0 and q=1. Likewise, from the spatial lags L for both the autocorrelation and partial autocorrelation functions we have identified $(\boldsymbol{\lambda}, \boldsymbol{\eta})$ to be $(0, 1_1)$. Hence, these both together would lead us to identify the linear orders $(p, q, \boldsymbol{\lambda}, \boldsymbol{\eta})$, i.e., we have $(p_{\boldsymbol{\lambda}}, q_{\boldsymbol{\eta}}) = (0, 1_1)$ giving

a STARMA $(0, 1_1)$ model. That is, there is the linear first order moving average with no autoregressive component, when assuming the model is a pure STARMA model.

We then estimated the values of the parameters $\boldsymbol{\psi} = (\theta_0, \theta_1)$ for this model using the procedures of Dai and Billard (2002). The resulting estimated values were $\hat{\boldsymbol{\psi}} = (-0.75, 0.15)$.

We then overfitted the model by adding extra parameters. Let the parameters of the general model be denoted by $\boldsymbol{\psi} = (\phi_0, \phi_1, \theta_0, \theta_1)$. A summary of the estimated parameter values along with the estimate of σ^2 for the residuals for some of these models is shown in Table 4.4.

For the so-called full model, we found $\hat{\boldsymbol{\psi}} = (\hat{\phi}_0, \hat{\phi}_1, \hat{\theta}_0, \hat{\theta}_1)$ = (-0.219, -0.118, -0.668, 0.247). From this, we conclude that the best model is STRAMA$(1_1, 1_1)$, the full model with mean square error $\hat{\sigma}^2 = 0.1349$. The Akaike Information criterion (AIC) $= nT\ log\hat{\sigma}^2 + 2$ (number of parameters) where n is the number of observations and T is the number of special sites, was calculated to be $AIC = 191(4)log(0.1349)+2(4) = -1,522.46$.

Hence, the estimated STARMA $(1_1, 1_1)$ model is

$$\hat{\mathbf{z}}(t) \;=\; \hat{\phi}_0\mathbf{z}(t-1) + \hat{\phi}_1\mathbf{W}^{(1)}\mathbf{z}(t-1) + \hat{\theta}_0\mathbf{e}(t-1) + \hat{\theta}_1\mathbf{W}^{(1)}\mathbf{e}(t-1) \qquad (4.4)$$

where $\mathbf{W}^{(1)}$ is our weighting matrix. Using the best-fit model (4.4), we then computed the predicted AIDS cases for each region and reverse transformed. The plot for both the reported and predicted values using the STARMA $(1_1, 1_1)$ model is given in Figures 4.3.a, 4.3.b, 4.3.c, and 4.3.d for the Northeast, Midwest, South, and East regions, respectively.

Comparing these plot of Figures 4.3.a with that in Figure 3.2.a for the ARMA model for the Northeast region, we can clearly see that the predicted values using the STRAMA model appear to give better fits than those obtained for the ARMA model, i.e., the predicted values depicted by broken lines in the STARMA plot

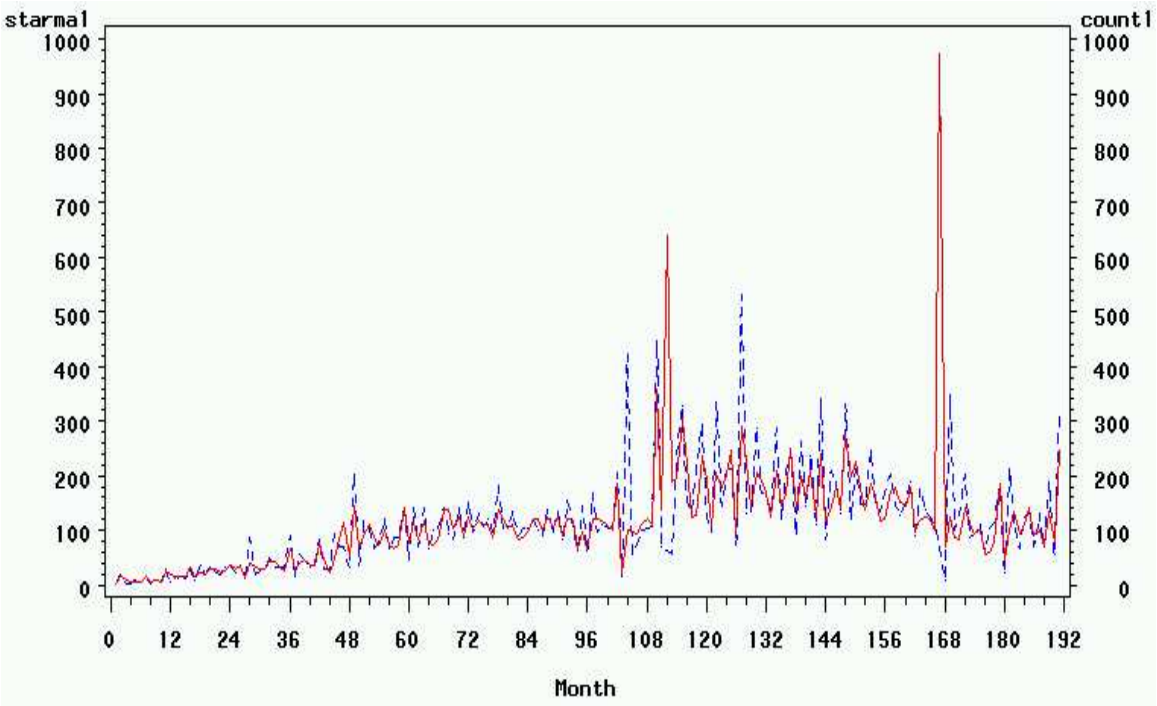


Figure 4.2.a Reported and Predicted AIDS Cases by Month, Northeast Region
Reported - Predicted ... using STARMA

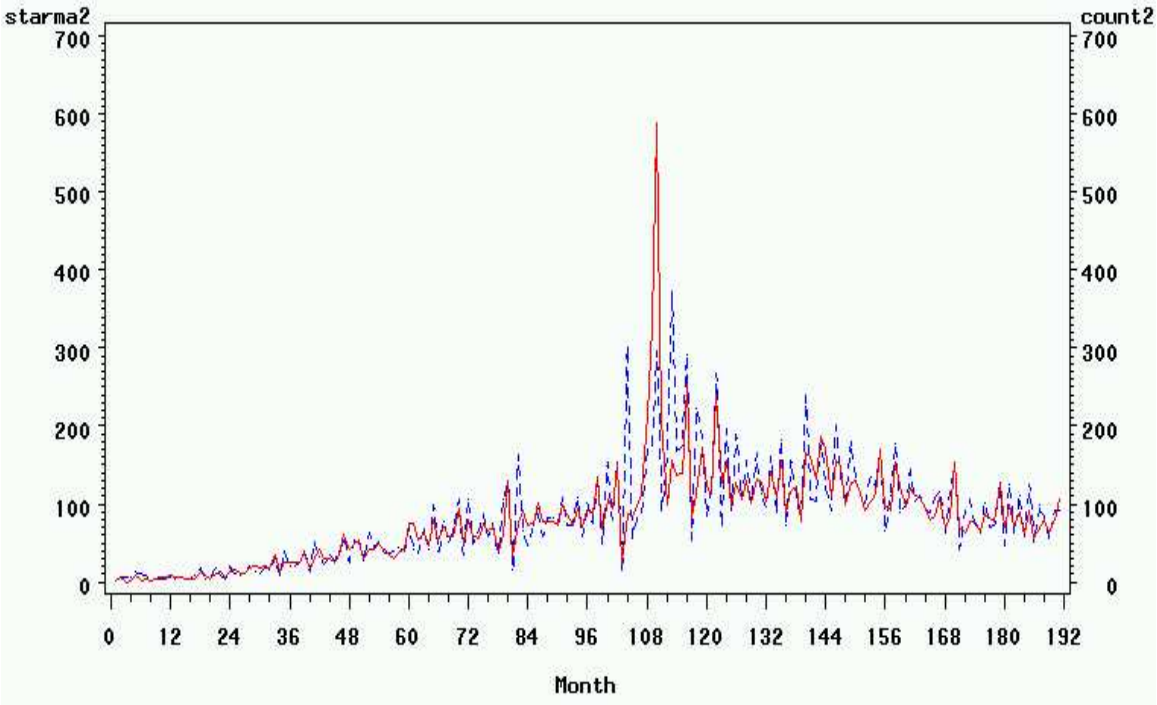


Figure 4.2.b Reported and Predicted AIDS Cases by Month, Midwest Region
Reported - Predicted ... using STARMA

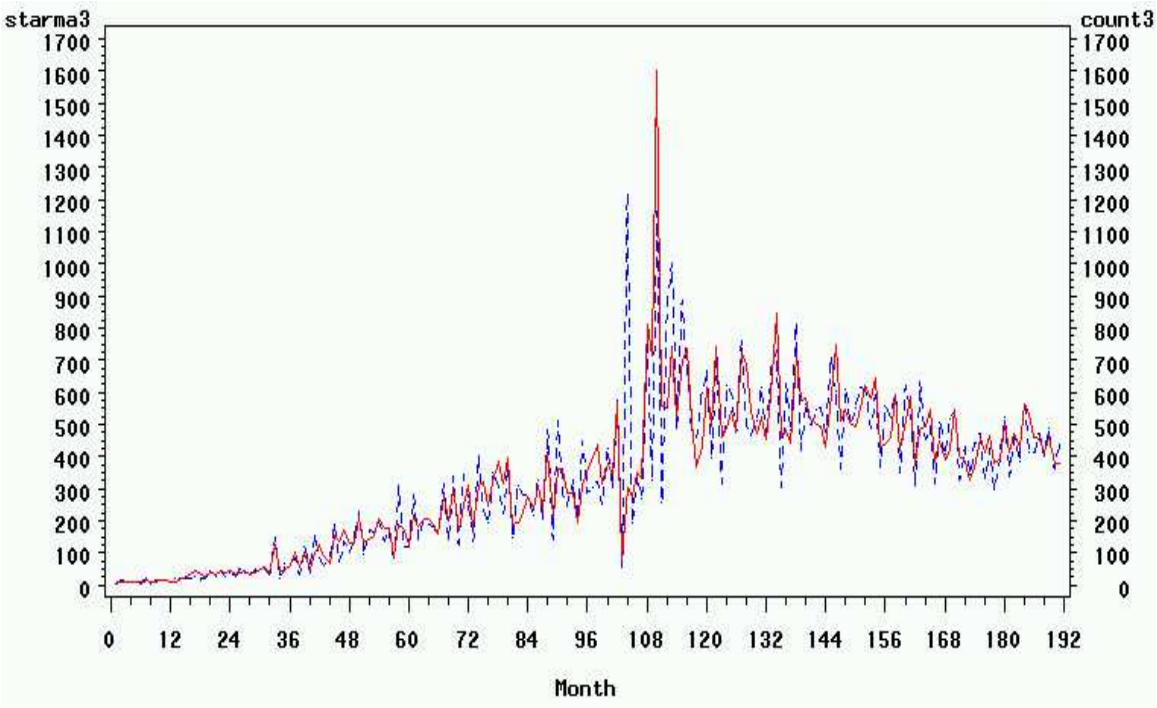


Figure 4.2.c Reported and Predicted AIDS Cases by Month, South Region
Reported - Predicted ... using STARMA

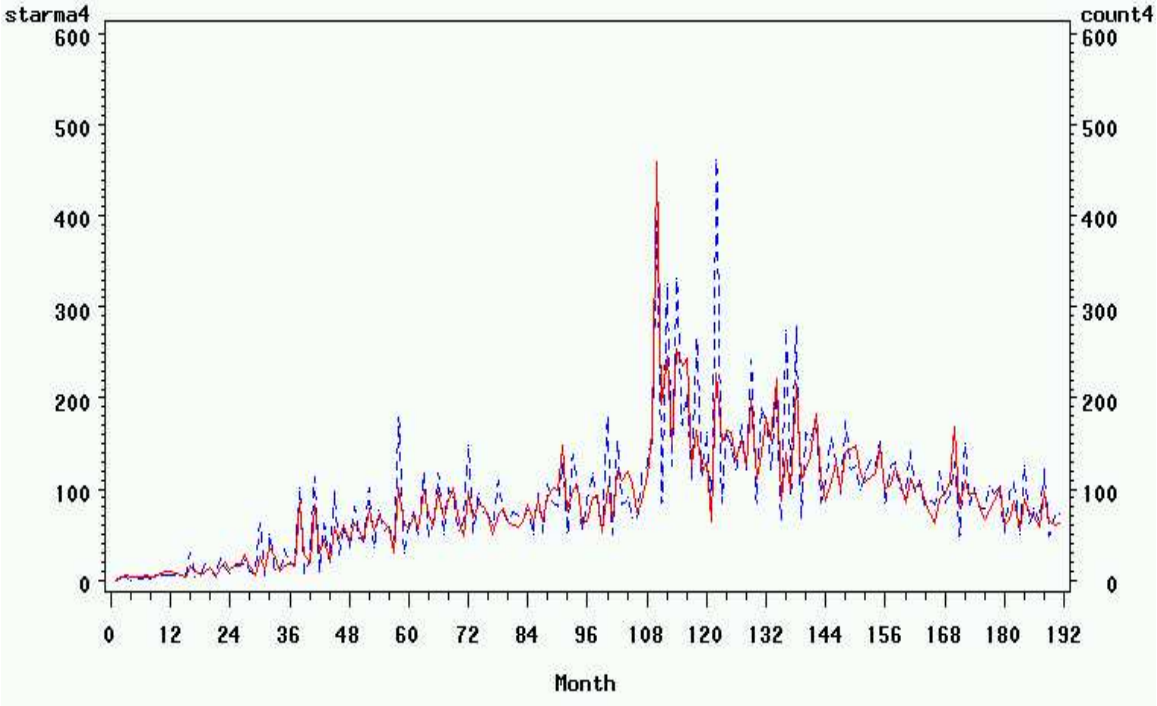


Figure 4.2.d Reported and Predicted AIDS Cases by Month, West Region
Reported - Predicted ... using STARMA

comes closer to the reported cases (solid lines). Likewise, plots comparing the observed and predicted number of cases for Region 2, Region 3 and Region 4 shows similar patterns. A fuller discussion of the comparison of these results with those obtained by fitting the nonspatial ARMA models is given in Chapter 5.

4.3 STBL Models

As for the previous models, in order to fit a STBL model to our data, it is first necessary to identify the model orders $(p, q, r, s, \boldsymbol{\lambda}, \boldsymbol{\eta}, \boldsymbol{\xi}, \boldsymbol{\mu})$. Dai and Billard (1998) developed a two-stage procedure to do this. The first stage involves finding the space-time autocorrelation function and the space-time partial autocorrelation function on the (transformed differenced) data, the $\mathbf{z}(t)$'s. This allows identification of the linear model orders $(p, q, \boldsymbol{\lambda}, \boldsymbol{\eta})$. This in effect corresponds to fitting a pure STARMA model. Thus, from the results of Section 4.2, we have identified $(p_{\boldsymbol{\lambda}}, q_{\boldsymbol{\eta}}) = (1_1, 1_1)$.

The second stage involves taking the residuals obtained after fitting the STARMA model identified in the first stage. These residuals then would correspond to a pure space-time bilinear model STBL$(0, 0, r_{\boldsymbol{\xi}}, s_{\boldsymbol{\mu}})$, i.e., the residuals contain only bilinear terms. From Dai and Billard, to identify the bilinear model orders $(r, s, \boldsymbol{\xi}, \boldsymbol{\mu})$ it is necessary to find the space-time autocorrelation on the residuals squared.

We have computed the autocorrelation and the partial autocorrelation functions using the squared residuals $\mathbf{e}^2(t)$ of the STARMA models, since the bilinear components of the STBL involves requires these $\mathbf{e}^2(t)$ as Dai and Billard discussed. That is, the input observations $\mathbf{z}(t)'s$ used in equation (4.2) and (4.3) are these $\mathbf{e}^2(t)'$s. The squared residuals $\mathbf{e}^2(t)'$s are the square of the difference between the reported and predicted values using the STARMA model. These autocorrelation

and the partial autocorrelation functions are shown in Table 4.5 and 4.6, respectively. Both the autocorrelation and the partial autocorrelation functions cut off at spatial lag L=1 while they do not exihibt a clear decay across the time lag. This reflects the presence of both the first order autoregressive and autoregressive and moving average terms in the model when assuming the model for the residuals is a pure STBL model.

Once again following the Dai and Billard maximum likelihood estimation procedure, we have estimated the values of the parameters for this model. Let us denote the vector of parameters as $\boldsymbol{\psi} = (\phi_0, \phi_1, \theta_0, \theta_1, \beta_{00}, \beta_{01}, \beta_{10}, \beta_{11})$.

Then, we found $(\hat{\phi}_0, \hat{\phi}_1, \hat{\theta}_0, \hat{\theta}_1, \hat{\beta}_{00}, \hat{\beta}_{01}, \hat{\beta}_{10}, \hat{\beta}_{11})$ = (-0.216, -0.073, -0.685, 0.204, 0.019, 0.0718, -0.118, 0.082) with $\hat{\sigma}^2 = 0.01341$, and hence $AIC = 191(4)log(0.1341) + 2(8) = -1,519.67$. Since $\hat{\phi}_1 = -0.073$, $\hat{\beta}_{00} = 0.019$ and $\hat{\beta}_{-1} = 0.07$ and $\beta_{11} = 0.082$ are very close to zero, a reduced model with parameters $\boldsymbol{\psi} = (\phi_0, \theta_0, \theta_1\beta_{10})$ was fitted. The resulting parameter estimates were $\hat{\sigma}^2 = 0.14$, with an AIC =-1,498.11. Comparing these results, we see that the full model is preferred.

Using the full model, we have forecasted the AIDS cases for each region and reverse transformed. The observed and predicted values are shown in Figures 4.3.a, 4.3.b., 4.3.c., and 4.3.d. for the Northeast, Midwest, South and West regions, respecitively. All plots shows best fit and reveal the usefulness of space time bilinear model that display shocks. In general, the bilinear models are better in modeling time series data which shows sudden outburst or (up and downs); compare Figures 3.2, 4.2 and 4.3. A fuller discussion of this comparison is presented in Chapter 5.

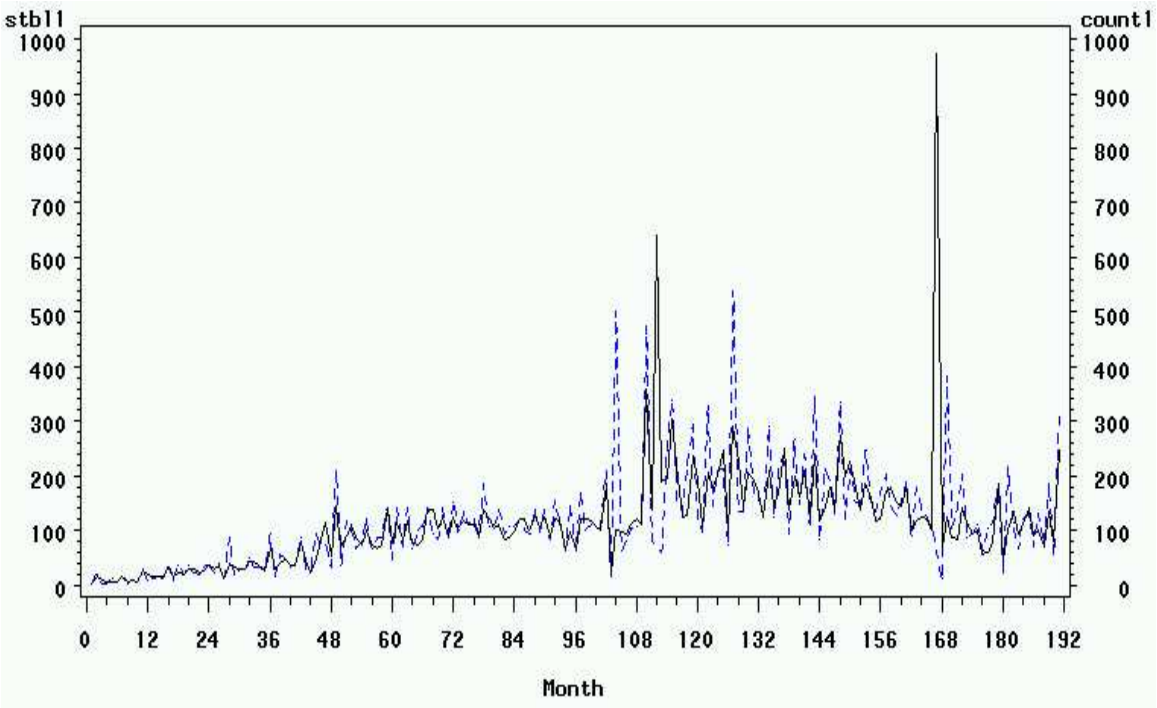


Figure 4.3.a Reported and Predicted AIDS Cases by Month, Northeast Region
Reported - Predicted ... using STBL

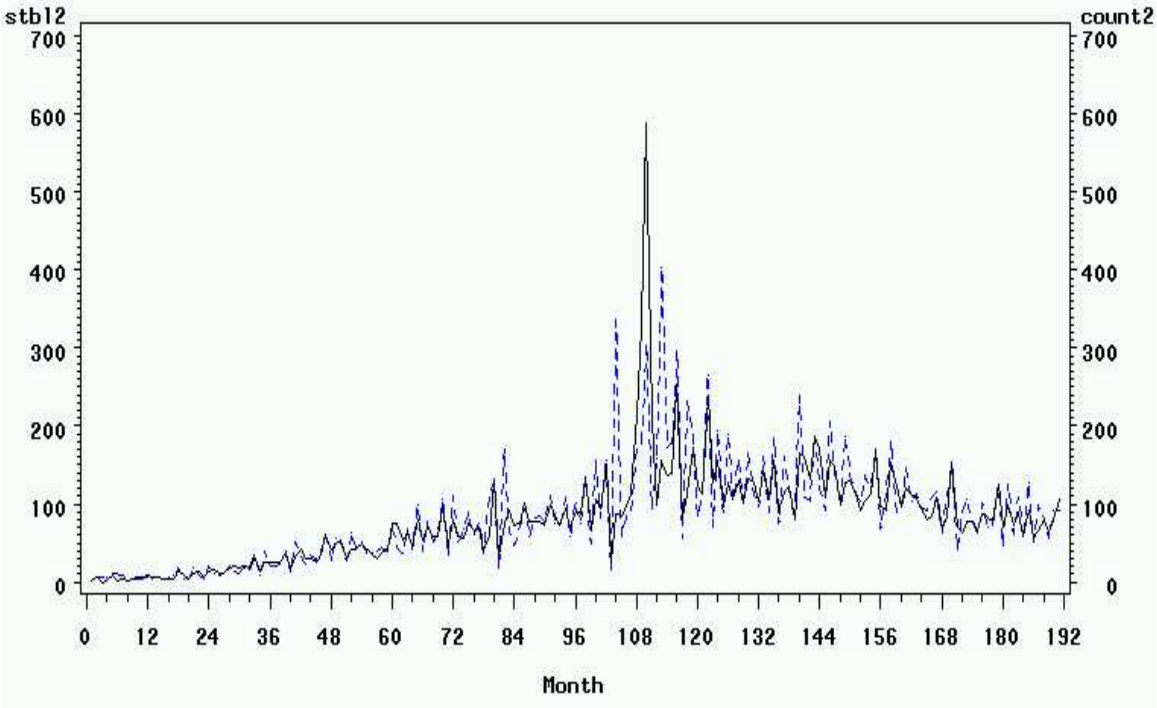


Figure 4.3.b Reported and Predicted AIDS Cases by Month, Midwest Region
Reported - Predicted ... using STBL

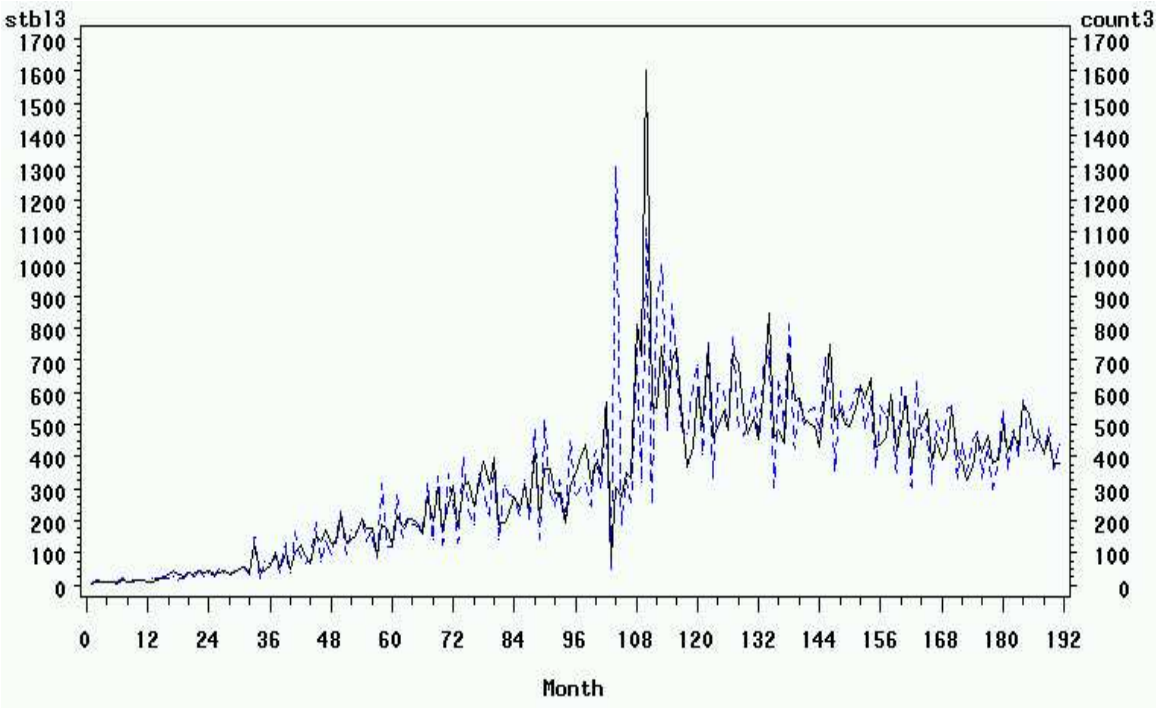


Figure 4.3.c Reported and Predicted AIDS Cases by Month, South Region
Reported - Predicted ... using STBL

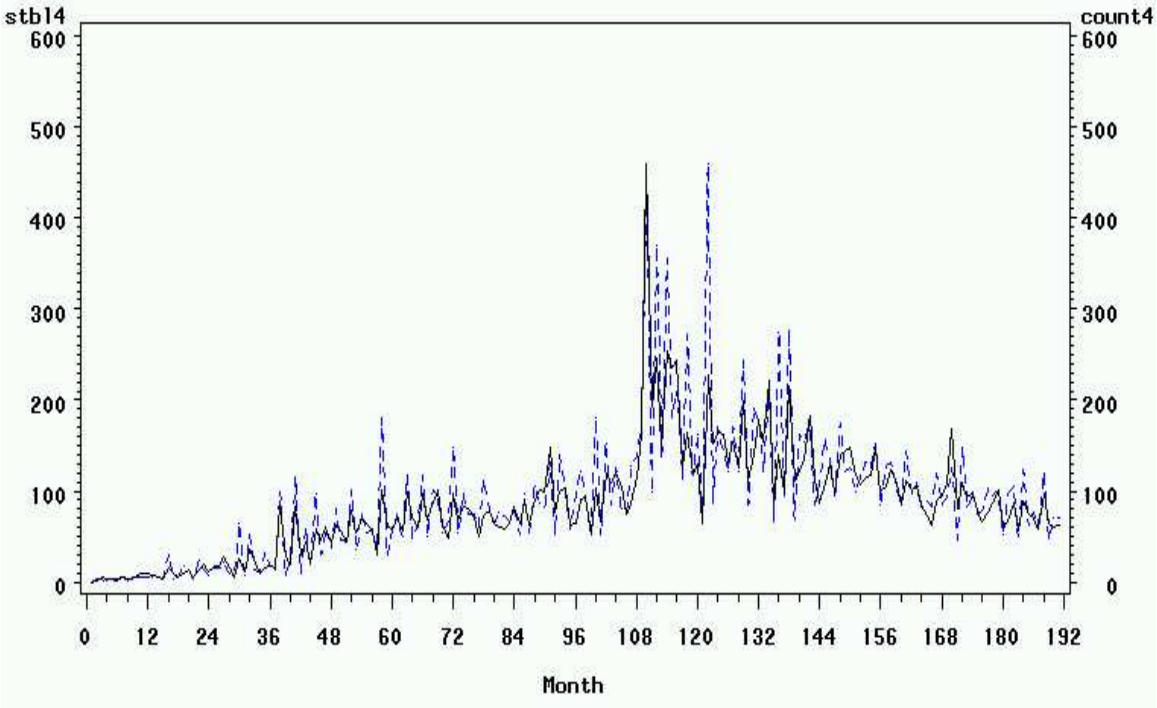


Figure 4.3.d Reported and Predicted AIDS Cases by Month, West Region
Reported - Predicted ... using STBL

Chapter 5

Comparisons and Conclusion

In Chapter 4, three different classes of models were fitted to our AIDS data. In this chapter, we compare these models. This is done by comparing estimated error variances $\hat{\sigma}^2$, and also by comparing the sum of squared forecast errors, for each model, for each region and in total over all regions.

For the estimated error variance $\hat{\sigma}^2$, we see that the respective ARMA $(0, 1)$ models gave estimates of $\hat{\sigma}^2$ equal to 0.177, 0.144, 0.095 and 0.142 for the Northeast, regions, respectively. When fitting the STARMA $(1_1, 1_1)$ model, this estimated value was $\hat{\sigma}^2$=0.1349; while for the STBL $(1_1, 1_1, 1_1, 1_1)$ model, the estimate was $\hat{\sigma}^2$=0.1341. Clearly, by this criterion, the STBL model provided the best fit and so would be preferred.

To compare the models by the sum of forecast error squares (SSF), we focussed attention on the twelve months January - December 1999. Table 5.1 summarizes this quantity for each of the three models by region and also gives the total SSF values over all regions.

When we examine the squared sum of forecast error (SSF) for the last twelve months, (January through December 1999) again the STBL model shows the best performance. The SSF for the univariate ARMA models, the full STARMA model and for the full STBL model are 209,755, 60,085 and 56,478, respectively. We see that also the STARMA model is a considerable improvement over the the standard univariate ARMA models. This means there is indeed spatial dependencies in the

spread of HIV/AIDS. However, the STBL model still has the best fit showing that in addition to the presence of spatial dependencies, there are indeed shocks that are not modeled by the STARMA model but which are captured by the STBL model. When we examine the SSF of each individual region, the STARMA model gives the best forecast for the Northeast, and Midwest regions (with very close margin), while the STBL shows the best fits for the South and West regions. This leads us to a close examination of the underlying process of our data, especially in the Northeast and South regions where the SSF values for the STARMA and STBL models show a wider range.

When we examine the Northeast region, we have observed one usually high monthly reported count of 974 surrounded by relatively low cases (with 103 average of the four surrounding counts). This surely looks like an outlier and will have a larger contribution to our the squared sum of forecast errors.

More importantly, the original time series AIDS counts for the South region fluctuates higher (up as high as 1,600 cases) with a standard deviation of 231.8 while the Northeast region counts exhibit relatively moderate variation (dance up and down slowly) with a standard deviation of 98.7 (compare Figure 4.2.a and Figure 4.3.c). This preference for the STBL model confirms its robust performance for those situations when the time series data exhibits such fluctuations. When we look at the plot of the STBL model (Figures 4.3) the original and the predicted data are very close (it would be a clear plot if the gap was wide), and we see that the STBL model is more sensitive to changes in the data.

The Space Time Bilinear (STBL) model gives the smallest $\hat{\sigma}^2$ and very close prediction for these AIDS data fitting, as expected. Comparison of Figures 3.2, 4.2, and 4.3 shows that the predicted values shown by broken lines in Figure 4.2 and 4.3 plots are very close to the reported cases (the solid lines) in some cases looks overlap which shows the closeness of the fit, while the ARMA model plots in

Figure 3.2 shows a wider gap. In general, both the STARMA and STBL models shows very competative best fit to our AIDS data.

This result is consistent with the findings of Dai and Billard with Mumps and Wind data. This reveals again that the STBL model has the potential to be applied to nonlinear spatial and temporal processes.

Chapter 6

Future Research

First, analogous to PROC ARIMA, develop PROC STARMA and PROC STBL a SAS friendly procedures by extending Dai and Billard's programs. This will make the application of bilinear and Space Time Bilinear models convenient for applied research activities.

Second, continue research on the application of space time bilinear models to many other areas of disciplines by giving emphasis on forecasting to confirm that STBL is a better time series model that can have a significant impact on quality and processing improvement.

Bibliography

Akaike, H. (1977), "On Entrophy Maximization Principle", *Proceedings Symposium on Applications of Statistics, Dayton Ohio, June 1977)* IP. R. Krishnaiah, ed.)

Aroian, L. A. (1980), Time Series in M-dimenstions: Definition problems, and Prospects. *Communications in Statistics* B9, 453-465.

Bickel, P. J. and Doksum, K. A. (1977), *Mathematical Statistics: Basic Ideas and Selected Topics",* Holden-day, Oakland.

Billard, L. and Dai, Y. (2000), "Modeling Spatial-Temporal Epidemics", *Technical Report,.*

Box, G. E. P. and Jenkins, G. M. (1976), *Time Series Analysis, Forecating and Control",* (2nd Edition), Holden-Day, San Franscisco.

Cliff, A. D., Haggett, P., Ord, J. K., Bassett, K. A. and Davies, R. G. (1975), *Elements of Statistical Structure: An Quantitative Approach,"* New York: Cambridge University Press.

Dai, Y. and Billard, L. (1998), "A Space Time Bilinear Model and its Identification", *Journal of Time Series Analysis,* 19, 657-679.

Dai, Y. and Billard, L. (2002), "Maximum Likelihood Estimation in Space Time Bilinear Models", *Journal of Time Series Analysis,* 23.

Durbin, J. and Koopman, S. J. (2001), "Time Series Analysis by State Space Methods", Oxford University Press, New York.

Gabr, M. M. (1993). "Maximum Likelihod Fitting of Bilinear Models to Time Series with Missing Observations", In *Developoments in Time Series Analysis,* (ed. T. Subba Rao), Chapman and Hall, UK, 283-291.

Gabr, M. M. and Subba Rao, T. (1981). "The Estimation and Prediction of Subset Bilinear Time Series Models with Applications", *Journal of Time Series Analysis,* 2, 155-71.

Granger, C. W. and Anderson, A. P. (1978), *"An Introduction to Bilinear Time Series Models"*, Vandenhoeck and Ruperct, Gottingen.

Harvey, A. C. (1989). "*Forecasting, Structural Time Series Models and the Kalman Filter,* " Cambridge University Press, Cambridge.

Harvey, A. C. (1981), "*Time Series Models*" New York: John Wiley.

Harvey, A. C. and Pierse, R. G. (1984), "Estimating Missing Observations in Economic Time Series", *Jounral of the American Statistical Association*, 79, 125-131.

Jones, R. H. (1980), "Maximum Likelihood Fitting of ARMA Models to Time Series with Missing Observations," *Techometrics*, 22, 389-395.

Kalman, R. E. (1960). "A New Approach to LInear Filtering and Prediction Problems", *Transactions ASME Journal of Basic Engineering* Series D, 83, 95-108.

Kim, W. K. and Billard, L. (1990), "Asymptotic Properties for the First-Order Bilinar Time Series Model", *Communications in Statistics Theory and Methods*, 19, 1171-83.

Kohn, R. and Ansley,C. F. (1983), "Exact Likelihood of Vector Autoregressive Moving Average Process with Missing or Aggregated Data", *Biometrika*, 70, 225-278.

Mohler, R. R. (1973), "*Bilinear control Processes,*" Academic Press, New York.

Oprian, C. Taneja, Voss, D. and Aroian, L. A. (1980), General Considerations and Interrelationships between MA and AR Models, Time Series in M-dimenstions, The ARMA Model. *Communications in Statistics,* B9, 515-532.

Pfeifer, P. E. and Deutsch, S. J. (1980a), "A Three-Stage Interative Procedure for Space-Time Modeling", *Technometrics*, 22, 25-47.

Pfeifer, P. E. and Deutsch, S. J. (1980b), "Identification and Interpretation of First Order Space-Time ARMA Models", *Technometrics,* 22, 297-408.

Sallas, W. M. and Harville, D. A. 91981). "Best Linear Recursive Estimation of Mixed Linear Model", *Journal of the American Statistical Association*, 76, 860-869.

Stehsholt, B. K. and Tjostheim, D. (1987), "Multiple Bilinear Time Series Models", *Journal of Time Series Analysis,* 8, 221-223.

Subba Rao, T. (1981), "On the Theory of Bilinear Models", *Journal of the Royal Statistics Society Series* B, 43, 244-255.

Subba Rao, T. and Gabr, M. M. (1984), "*An Introduction to Bispectral Analysis and Bilinear Time Series Models,*" Springer-Verlag, Berlin.

Taneja, V. S. and Aroian, L. A. (1980), Time Series in M-dimenstions: Autoregressive Models. *Communications in Statistics,* B9, 491-513.

Voss, D. A. Oprian, C. A. and Aroian, L. A. (1980), "Moving Average Models-Time Series in M Dimensions", *Communications in Statistics,* B9, 467-489.

Appendix A

Tables

A.1 Table 3.1. U.S. Census Bureau Regions Division

Region	State		
Northeast	Connecticut	New Jersey	Maine
	New York	Massachusetts	Pennsylvania
	New Hampshire	Rhode Island	Vermont
Midwest	Indiana	Iowa	Nebraska
	Illinois	Kansas	North Dakota
	Michigan	Minnesota	South Dakota
	Ohio	Missouri	Wisconsin
South	Delaware	Alabama	Arkansas
	District of Columbia	Kentucky	Louisiana
	Florida	Mississippi	Oklahoma
	Georgia	Tennessee	Texas
	Maryland	North Carolina	South Carolina
	Virginia	West Virginia	
East	Arizona	Montana	Alaska
	Colarado	Utah	California
	Idaho	Nevada	Hawaii
	New Mexico	Wyoming	Oregon
	Washington		

A.2 Table 3.2. The First Order Neighborhood Structure

Region Code	Region	Neighbor Regions
1	Northeast	Midwest and South
2	Midwest	Northeast, South and West
3	South	Northeast, Midwest and West
4	West	Midwest and South

A.3 Table 4.1. Summary of the Estimation of ARIMA Procedure

$\hat{\phi}_0$	$\hat{\phi}_1$	Region	$\hat{\sigma}^2$	AIC
0.013	0.772	Northeast	0.177	213.634
0.019	0.721	Midwest	0.144	174.078
0.019	0.673	South	0.095	93.944
0.017	0.722	West	0.142	171.942

A.4 Table 4.2. Space-Time Autocorrelation for $\{Z(t)\}$, STARMA Model

Time lag	Spatial lag L = 0	spatial lag L = 1
S=1	-.5403	-.2413
S=2	0.0539	0.0998
S=3	0.0292	-.0188
S=4	-.0054	-.0090
S=5	-.03157	0.0075
S=6	0.0491	0.0499
S=7	-.0515	-.1139

A.5 Table 4.3. Space-Time Partial Autocorrelation for $\{Z(t)\}$, STARMA Model

Time lag	Spatial lag L = 0	Spatial lag L = 1
S=1	-.5403	-.0541
S=2	-.3367	0.0805
S=3	-.1789	0.1066
S=4	-.0925	0.0719
S=5	-.0961	0.0894
S=6	-.0271	0.1818
S=7	-.0685	-.0082

A.6 Table 4.4. Summary for Estimation of STARMA Model

Model	$\hat{\phi}_0$	$\hat{\phi}_1$	$\hat{\theta}_0$	$\hat{\theta}_1$	$\hat{\sigma}^2$	AIC
STARMA $(0, 1_1)$	0	0	- 0.75	0.15	0.1408	-1,493.76
STARMA $(1_1, 0)$	-0.52	-0.04	0	0	0.1605	-1,393.71
STARMA $(1_1, 1_1)$	-0.219	-0.118	-0.668	0.247	0.1349	-1,522.46

A.7 Table 4.5. Space-Time Autocorrelations for $\{Z^2(t)\}$, STBL Model

Time lag	Spatial lag L=0	Spatial lag L=1
S=1	0.0041	0.0001
S=2	-.0054	0.0005
S=3	-.0058	0.0002
S=4	-.0067	-.0003
S=5	0.0248	0.0002
S=6	0.0467	0.0049
S=7	0.0047	-.0001

A.8 Table 4.6. Space-Time Partial Autocorrelations for $\{Z^2(t)\}$, STBL Model

Time lag	Spatial lag L = 0	Spatial lag L = 1
S=1	-.0041	0.0001
S=2	-.0054	0.0005
S=3	-.0059	0.0003
S=4	-.0068	-.0002
S=5	0.0247	0.0002
S=6	0.0468	0.0032
S=7	0.0054	-.0001

A.9 Table 5.1. Reported and STBL Predicted (in parenthesis) AIDS Cases in 1999

Month	Northeast	Midwest	South	East
Jan	47(23)	71(48)	501(537)	61(53)
Feb	100(214)	98(126)	412(340)	66(97)
Mar	134(115)	72(62)	472(479)	87(105)
Apr	93(65)	90(109)	433(399)	56(50)
May	111(122)	61(58)	563(577)	89(123)
Jun	142(131)	91(126)	532(421)	74(62)
Jul	89(70)	55(51)	457(417)	71(77)
Aug	101(130)	67(98)	453(481)	58(63)
Sep	71(73)	79(83)	411(409)	97(120)
Oct	135(187)	63(55)	458(491)	65(47)
Nov	79(55)	81(91)	378(363)	61(70)
Dec	245(309)	106(91)	380(436)	62(71)

A.10 Table 5.2. Summary of Forcast Error Square in 1999

Region	ARMA	STARMA	STBL
Northeast	27561.19	23048.59	23592.77
Midwest	24328.06	4478.35	4495.71
South	131696.94	28596.49	24694.89
East	26169.44	3962.21	3695.09
Total	209755.63	60085.64	56478.46

Appendix B

DATA SETS AND PROGRAMS

B.1 DATA SETS

The data set for this paper contains the year and month in which CDC received the AIDS case report, and region (non-metropolitan area) of residence at diagnosis of AIDS identified since January 1984 through December 1999 (refer to section 3.1). Below is the partial list of the data file (Region 1=Northeast, 2=Midwest, 3=South and 4=West).

Region	Year	Month
1	84	01
1	84	02
1	84	03
more		
4	99	10
4	99	11
4	99	12

B.2 SAS PROGRAMS

```
/*************************************************************
*  Program:    aids_pgm.sas
*  Programmer: Tadesse Haileyesus, April 2002
*  Input:      CDC External ASCII file called pids.txt
*  Desc:       To Apply Time Series Models to AIDS Data
*************************************************************/
options ls=78 ps=56 nodate formdlim='-' pageno=1;
data temp; set maps.us;
/* set 'C:\Program Files\SAS Institute\SAS\V8\maps.us'; */
if state in('9','23','25','33','44','50','34','36','42')
then DIVISION = 1 ;
else if state in('18','17','26','39','55','19',
                 '20','27','29','31','38','46')
then DIVISION = 2 ;
else if state in('10','11','12','13','24','37','45','51',
                 '54','1','21','28','47','5','22','40','48')
then DIVISION = 3 ;
else if state in('4','8','16','35','30','49','32',
                 '56','2','6','15','41','53')
then DIVISION = 4 ;
else DIVISION = 5;
run;

proc sort data=temp;
by division;
run;
proc gremove data=temp out=remstate;
   by division;
   id state;
run;
goptions reset=global gunit=pct
         border=white cback=white colors=(black )
         ftext=swiss htitle=6 htext=3;
footnote 'Figure 3.1. U.S. Census Bureau Regions Division Map';
proc gmap map=remstate data=temp all;
   id division;
   choro division / discrete legend=legend ; * nolegend;
run;
quit;
```

```
/* Import year, month and region from CDC AIDS data */
data data1;
  infile 'N:\aids\pids.txt';
  input year 3-4 month 5-6 region 10;
run; /* DATA1 has 115,482 observations and 3 variables, 2.7 MB */

/* Group by region, year, month to get counts */
proc sql;
  create table data2 as
    select region, year, month, count(*) as count
    from data1 where year >= 84  /* get Jan 84 to Dec 99 */
    group by region, year, month;
/* DATA2 created, with 765 rows and 4 columns. */

data time; /* create time frame */
  do region = 1 to 4;
    do year = 84 to 99;
      do month = 1 to 12;
         time + 1; /* 16yrs x 12=192 monthly reports */
         If time > 192 then time=1; output;
      end;
    end;
  end;
run; /* TIME has 768 observations and 4 variables. */

proc sort data=time;
  by region year month;
run;

data merge1;
  merge time data2;
by region year month;
run; /* MERGE1 has 768 obs and 5 var */

/* handle missing values as discussed in section 3.1 */
data data3;
  set merge1;
  if (region =4 and year =92 and month =8) then count=110;
  if count =. then count=0;
put count@@;
run;
```

```
/* set date, transform data and output */
data aids.trans_tbl; /* out put table */
  set data3;
    file 'N:\aids\trans_file'; /* out put file */
    date =mdy(month,1,year);
    format date monyy.;
  sqct=sqrt(count);
  logct=log(1+count);
  put region year month count sqct logct;
run;

/* out put in separate files: region1 Northeast (shown below)
region2 Midwest, region3 South and region4 West (not shown)  */
data region1 ;
  set data4;
    file 'N:\aids\region1';
    if region=1 then
    put region year month count sqct logct;
run;

/* To do the identification stage for log(1+count)*/
Title 'Figure 3.3 Partial & Autocorrelation Functions,
        Northeast Log Transformed Data';
PROC ARIMA data=aids.region1 out=out1;
  Identify var=logct nlag=48;

Title ' Figure 4.1 Partial & Autocorrelation Functions,
  Northeast Log Transformed and First Differenced Data';
PROC ARIMA data=aids.region1 out=out1a;
  Identify var=logct(1) nlag=48;

/* To get first difference, AIDS data */
/* output count1234, d1sqct1234 & d1logct1234, 191 obs, 4 regions*/

data reg1 (keep=time count1 sqct1 logct1);
  infile 'N:\aids\region1';
  input time region year month count1 sqct1 logct1 ;

data aids.region1234;
  merge reg1 reg2 reg3 reg4;
  by time; /* region1234 merged side by side */
```

```
data aids.d1region1234
(keep= dcount1 dsqct1 dlogct1 dcount2 dsqct2 dlogct2
       dcount3 dsqct3 dlogct3 dcount4 dsqct4 dlogct4);
  set aids.region1234 ;
  dcount1 = dif1(count1);
  dsqct1 = dif1(sqct1);
  dlogct1 = dif1(logct1);

  dcount2 = dif1(count2);
  dsqct2 = dif1(sqct2);
  dlogct2 = dif1(logct2);

  dcount3 = dif1(count3);
  dsqct3 = dif1(sqct3);
  dlogct3 = dif1(logct3);

  dcount4 = dif1(count4);
  dsqct4 = dif1(sqct4);
  dlogct4 = dif1(logct4);

data aids.d1count1234 (keep=dcount1 dcount2 dcount3 dcount4 );
  set aids.d1region1234 ;
  where dcount1 <> . ;
data aids.d1sqct1234 (keep=dsqct1 dsqct2 dsqct3 dsqct4 );
  set aids.d1region1234 ;
  where dsqct1 <> . ;
data aids.d1logct1234 (keep=dlogct1 dlogct2 dlogct3 dlogct4 );
  set aids.d1region1234 ;
  where dlogct1 <> . ;

data in;
  set aids.d1logct1234;
  file 'N:\aids\d1logct1234';/*transformed & diff1 out put file*/
  put dlogct1-dlogct4;
run;

/* Do estimation, forecast and reverse transform */
/* We work with the (0,1,1) of the (p,d,q)x(P,D,Q)
for s=1 model, as this gave best fit */

PROC ARIMA data=aids.region1  out=outt1;
  Identify var=logct(1) nlag=48 noprint outcov=outcvl1;
  Title2 'Log Counts --- First Difference';
```

```
  estimate  q=(1) grid plot; run;
  FORECAST back=12 lead=12 interval=month out=flog1; run;
quit;
/* To put data outcvl1 into an outfile <ocvl1> */
data regl1; set outcvl1;
  file 'N:\aids\ocvl1';
  code=1;
  put LAG 6-7 VAR 10-13 N 16-18 COV 10.5 CORR 10.5 STDERR 10.5
      INVCORR 10.5 PARTCORR 10.5 code 2-3; run;
proc print data=regl1; run;

/* To put forcast into outfile <forsl1> */
data data1; set flog1;
  file 'N:\aids\forl1';
  code=1;
  put logct 12.8 forecast 12.8 std 12.8 l95 12.8 u95 12.8
      residual 12.8 code 5-6 ; run;
proc print data=data1; run;

/* To inverse transform forcast */
data aids.arma1; set flog1;
  count1=-1 + exp(logct);
  ll95=-1 + exp(l95);
  lu95=-1 + exp(u95);
  arma1=-1 + exp( forecast + std*std/2);
  lresid=count1-arma1; run;
run; /* similar steps for region 2,3 & 4 */

data test1 (keep= arma1 count1);
 set aids.arma1; if _N_ > 1 ;
data test1; set test1 ; Month+1;
run;

data aids.arma_final;
merge test1 test2 test3 test4 ;
by Month;
run;

/* ARMA PLOT */
proc gplot data =aids.arma_final;
  plot arma1*Month/haxis = 0 to 192 by 12
                   vaxis = 0 to 1000 by 100;
  plot2 count1*Month/vaxis = 0 to 1000 by 100;
```

```
  symbol1 value=point color=blue line=20 i=join;
  symbol2 value=none color=red line=1 i=join;
  footnote1'Figure 3.2.a Reported and Predicted AIDS Cases,
  Northeast Region' height=3;
  footnote2 '-, reported; ---, predicted using ARMA';
run;

/* STARMA correlations */
Title1 'U.S. (non MSA) AIDS Cases, First Difference';
%let slag=1;
%let tlag=12;

data weights;                /* 0.0 0.2 0.8 0.0 */
infile 'N:\aids\PPSW.data';  /* 0.2 0.0 0.6 0.2 */
input w1-w4 ;                /* 0.4 0.3 0.0 0.3 */
run; /* 4x4 */               /* 0.0 0.2 0.8 0.0 */

data obsern;
set aids.d1logct1234;
run;  /* 180x4 */

proc iml ;
use weights;
read all into w;

use obsern;
read all into dat;
obs=dat`; /* 4x180 */

/*if univariate bilinear series, set w={0} */
n=nrow(obs);
t=ncol(obs);
w=i(n)||w;
max_obs=max(obs);
min_obs=min(obs);
print min_obs max_obs n t w ;
     /*  -861    1165 4 180 */

start cov(n, t, w, z, var);
sg=j(&slag+1, &slag+1, 0);

cn="L=0":"L=&slag";
rn="S=1":"S=&tlag";
```

```
bg=covlag(z', &tlag+1);
do s=1 to &tlag+1;       /* s=time lag*/
do l=1 to &slag+1;       /*l=space lag*/
do k=1 to &slag+1;
wwg=t(bg[,(s-1)*n+1:s*n]);

wwg=w[,(k-1)*n+1:k*n]'*w[,(l-1)*n+1:l*n]*wwg;
sg[l,k]=trace(wwg)/n;
end;
end;
if s=1 then msg=sg;
 else msg=msg//sg;
end;

msg=toeplitz(msg);  /*this is covariance matrix*/
dmsg=vecdiag(msg[1:(&slag+1),1:(&slag+1)]);

do s=1 to &tlag;
acf=msg[s*(&slag+1)+1:(s+1)*(&slag+1),1];
acf=acf/sqrt(dmsg#msg[1]);
if s=1 then lo_l0_s=acf';
 else lo_l0_s=lo_l0_s//acf';
end;
print var;
print 'Space-Time Autocorrelations (L-Spatial Lag, S-Time Lag)';
print lo_l0_s[rowname=rn colname=cn format=6.4];

/*.....PACF....*/
u=0;
phi_sl=j(&tlag,&slag+1,0);
do s=1 to &tlag;
do l=1 to &slag+1;
u=u+1;
a=msg[1:u,1:u];
b=msg[&slag+2:&slag+1+u,1];
pacf=solve(a,b);
phi_sl[s,l]=pacf[u];
end;
end;
print 'Space-Time Partial Autocorrelations (L-Spatial Lag,
 S-Time Lag)';
print phi_sl[rowname=rn colname=cn format=6.4];
```

```
finish;
zsquare=obs##2;
run cov(n,t,w,zsquare,"Z(t) Square");

run cov(n,t,w,obs,"Z(t)");
quit;

/* fit STARMA */
title 'Predicted AIDS, log ct diff1 STARMA(1[1], 1[1]) Model';

data realdata;
set aids.Count1234 ; /* 192 original data */
if _N_ >1;           /* 191 obs */
output; run;

data diffdata;
set aids.d1logct1234 ; /* 191 log transformed & first diff data */
output; run;

data weights;
infile 'N:\aids\PPSW.data';
input w1-w4 ;
run;
title ;

proc iml symsize=900000;
use weights;
read all into w;

use diffdata;
read all into dlogct;
z=dlogct';

/* input known parameter values */
para={-0.218668, -0.117796, -0.667915, 0.2471533, 0, 0, 0, 0};
* print para;

nobs= ncol(z);
g=ncol(w);
zs=j(g,1,0);
pz=j(g,nobs,0);
e=j(g,nobs,0);
predz=j(g,1,0);
```

```
do t=2 to nobs;
bb=w*z[,t-1];
ee=w*e[,t-1];
y=z[,t-1]||bb||e[,t-1]||ee||z[,t-1]#e[,t-1]||
    z[,t-1]#ee||bb#e[,t-1]||bb#ee;
free ee bb;
zs=y*para;
e[,t]=z[,t]-zs;
npredz=zs;
predz=predz||npredz;
if t=nobs then do;
STARMA=predz‘;
eout=e‘;

*print STARMA; /* 191 predz data */
print eout;
end; end;
run;
quit;

data one;
set aids.logct1234 ; /* 192-1 log transformed */
 if _N_ > 1;
proc print; run;
data one; set one;
t=_N_;       /* 191 obs */
output;
proc print; run;
proc sort data=one;
   by t;
data two;
  infile 'N:\aids\STARMA_predicted.txt'; /* 191 predicted z data */
  input forcast1-forcast4;
  t=_N_;
output;
proc print; run;
proc sort data=two;
   by t;


data aids.starma ;*(keep=t starma1 starma2 starma3 starma4 );
merge one two;
```

```
by t;           /* 191-1 mse */
starma1=-1+exp(logct1+forcast1);
starma2=-1+exp(logct2+forcast2);
starma3=-1+exp(logct3+forcast3);
starma4=-1+exp(logct4+forcast4);
output;
proc print; run;

data three;
set aids.Count1234 ; /* 192-1 original count */
if _N_ > 1;
data three; set three ;
t=_N_;
data aids.starma_final ;*(keep=t starma1 starma2 starma3 starma4 );
merge aids.starma three;
by t;           /* 191-1 mse */
options ls=200 ps=230;
proc print; run;

/* STBL correlations */
Title1 'U.S. AIDS Cases, First Difference for STBL using res squares';
%let slag=1;
%let tlag=12;

data weights;                /* 0.0 0.2 0.8 0.0 */
infile 'N:\aids\PPSW.data';  /* 0.2 0.0 0.6 0.2 */
input w1-w4 ;                /* 0.4 0.3 0.0 0.3 */
run; /* 4x4 */               /* 0.0 0.2 0.8 0.0 */

data obsern;
set aids.Starma_ee; /* using starma error/residual squares */
run;  /* 191x4 */

proc iml ;
use weights;
read all into w;

use obsern;
read all into dat;
obs=dat`; /* 4x191 */

/*if univariate bilinear series, set w={0} */
n=nrow(obs);
```

```
t=ncol(obs);
w=i(n)||w;
max_obs=max(obs);
min_obs=min(obs);
print min_obs max_obs n t w ;
     /*  -861    1165 4 191 */

start cov(n, t, w, z, var);
sg=j(&slag+1, &slag+1, 0);

cn="L=0":"L=&slag";
rn="S=1":"S=&tlag";

bg=covlag(z‘, &tlag+1);
do s=1 to &tlag+1;       /* s=time lag*/
do l=1 to &slag+1;       /*l=space lag*/
do k=1 to &slag+1;
wwg=t(bg[,(s-1)*n+1:s*n]);

wwg=w[,(k-1)*n+1:k*n]‘*w[,(l-1)*n+1:l*n]*wwg;
sg[l,k]=trace(wwg)/n;
end;
end;
if s=1 then msg=sg;
 else msg=msg//sg;
end;

msg=toeplitz(msg);  /*this is covariance matrix*/
dmsg=vecdiag(msg[1:(&slag+1),1:(&slag+1)]);

do s=1 to &tlag;
acf=msg[s*(&slag+1)+1:(s+1)*(&slag+1),1];
acf=acf/sqrt(dmsg#msg[1]);
if s=1 then lo_l0_s=acf‘;
 else lo_l0_s=lo_l0_s//acf‘;
end;
print var;
print ’Space-Time Autocorrelations (L-Spatial Lag, S-Time Lag)’;
print lo_l0_s[rowname=rn colname=cn format=6.4];

/*.....PACF....*/
u=0;
phi_sl=j(&tlag,&slag+1,0);
```

```
do s=1 to &tlag;
do l=1 to &slag+1;
u=u+1;
a=msg[1:u,1:u];
b=msg[&slag+2:&slag+1+u,1];
pacf=solve(a,b);
phi_sl[s,l]=pacf[u];
end;
end;
print 'Space-Time Partial Autocorrelations (L-Spatial Lag,
 S-Time Lag)';
print phi_sl[rowname=rn colname=cn format=6.4];

finish;

zsquare=obs##2;
run cov(n,t,w,zsquare,"Z(t) Square");

run cov(n,t,w,obs,"Z(t)");
quit;

/* STBL fit */
data realdata;
set aids.Count1234 ; /* 192 original data */
if _N_ >1;           /* 191 obs */
output; run;

data diffdata;
set aids.d1logct1234 ; /* 191 log transformed & first diff data */
output; run;

data weights;
infile 'N:\aids\PPSW.data';
input w1-w4 ;
run;
title ;

proc iml symsize=900000;
use weights;
read all into w;

use diffdata;
read all into dlogct;
```

```
z=dlogct‘;

/* input known parameter values */
para={-0.218668, -0.117796, -0.667915, 0.2471533,
     0.0190881, 0.0710594, -0.117627, 0.0824297};

nobs= ncol(z);
g=ncol(w);
zs=j(g,1,0);
pz=j(g,nobs,0);
e=j(g,nobs,0);
predz=j(g,1,0);

do t=2 to nobs;
bb=w*z[,t-1];
ee=w*e[,t-1];
y=z[,t-1]||bb||e[,t-1]||ee||z[,t-1]#e[,t-1]||
    z[,t-1]#ee||bb#e[,t-1]||bb#ee;
free ee bb;
zs=y*para;
e[,t]=z[,t]-zs;
npredz=zs;
predz=predz||npredz;
if t=nobs then do;
STBL=predz‘;
eout=e‘;

print STBL; /* 191 predz data */
*print eout;
end; end;
run;
quit;

data one;
set aids.logct1234 ; /* 192-1 log transformed */
 if _N_ > 1;
proc print; run;
data one; set one;
t=_N_;       /* 191 obs */
output;
proc print; run;
proc sort data=one;
   by t;
```

```
data two;
  infile 'N:\aids\STBL_predicted.txt'; /* 191 predicted z data */
  input forcast1-forcast4;
  t=_N_;
output;
proc print; run;
proc sort data=two;
   by t;

data aids.stbl ;
merge one two;
by t;           /* 191-1 mse */
stbl1=-1+exp(logct1+forcast1);
stbl2=-1+exp(logct2+forcast2);
stbl3=-1+exp(logct3+forcast3);
stbl4=-1+exp(logct4+forcast4);
output;
proc print; run;

data three;
set aids.Count1234 ; /* 192-1 original count */
if _N_ > 1;
data three; set three ;
t=_N_;
data aids.stbl_final ;
merge aids.stbl three;
by t;           /* 191-1 mse */
options ls=200 ps=230;
run;
```