ARIMA MODELING OF TRAFFIC FLOW DATA

John S. Pipkin
Department of Geography and Planning
State University of New York at Albany
Albany, NY 12222

ABSTRACT Geographers have devoted little attention to modeling temporal variability in traffic flow at daily or hourly resolutions. ARIMA (Auto-Regressive Integrated Moving Average) models are suitable for this purpose. The Data Services Bureau of New York State Department of Transportation maintains hourly, directional traffic volume counts for about 60 continuous recording stations on state-funded highways throughout New York. The counts have been recorded at some sites since the mid-1970's. Although equipment malfunctions and other problems produce many "missing values," the data still comprise a uniquely detailed record of traffic flow characteristics in a variety of geographic settings in the state. General characteristics of the data are outlined. The treatment of missing values poses special problems. Sites typically fall into one of two-classes: those dominated by the worktrip, in which daily volumes are highest and most predictable on weekdays; and those dominated by discretionary traffic. The latter show the greatest seasonal variability; their flow usually peaks at weekends. Problems of time series modeling are discussed. Two types of seasonality dominate the data: day-of-week effects and seasonality by time of year. The basic features of seasonal ARIMA modeling are outlined. Previous day and previous week (seasonal) influences are divided into integration effects (trend or drift), autoregressive effects (effects of previous series values), and moving average effects (effects of previous random shocks or error terms). With appropriate model specification and identification, ARIMA provides a coherent representation of all these influences, which can be interpreted in terms of substantive traffic characteristics at various sites. The overall fit of ARIMA models to the data is good. A typical model sequence comprises: a logarithmic transformation to attain homoscedasticity; first-order differencing (both regular and seasonal) to attain stationarity; and first order autoregressive and moving average effects (regular and seasonal). The residual series can be applied to detect special traffic events, including the effects of winter snowstorms. Examples are given from highways in the Capital District of Albany, New York.

INTRODUCTION

This paper outlines an application of ARIMA (Autoregressive Integrated Moving Average) models to traffic flow data. The work formed part of a larger effort to examine the effects of weather on traffic flow in New York state, using a large data set on traffic available from New York State Department of Transportation.[1] The broader project examines the extent to which weather events such as winter snow storms and warm weekends in spring produce discernible short term variations in traffic flow, in discretionary trips (such as shopping and recreation) and in the non-discretionary work trip. The general strategy for detecting such fluctuations has been to examine the residuals from various

[1]For access to the data and for several helpful discussions the assistance of the Data Services Bureau, New York State Department of Transportation, is gratefully acknowledged. The data are described in *Regional Traffic Volume Data from Continuous Count Stations in New York State: 1972-1982* (mimeo), Transportation Statistics and Analysis Section, Planning Division Data Services Bureau, New York State Department of Transportation, Albany, NY 12232, July 1983. Many helpful discussions with John T. Hayes are also gratefully acknowledged.

time-series models of traffic flow, including ARIMA models, moving averages and several types of polynomial regression. In principle, after day-of-the-week and seasonal effects have been removed and an appropriate model of residual variation has been fitted, weather related events may be detected by comparing flow residuals with known series of weather records. Such modeling experiments may contribute to our general understanding of the geography of traffic patterns. They may also yield models useful in assessing (and mitigating) the direct and indirect effects of winter storms on travel in the snowy northeast. This work is has been presented more fully elsewhere.[2]

The objective in this paper is to outline the ARIMA technique and to show its applicability to traffic flow series per se (without any reference to climatic series). This seems worthwhile because ARIMA has not been widely applied in transportation geography and because the fits obtained on the NYS DOT traffic data are extraordinarily good.

DATA

The data comprise hourly, directional traffic volume counts for 59 continuous recording stations on statefunded highways in New York. The counts have been recorded at some sites since the mid-1970's. We have examined in detail data for 1980, 1981, 1986 and 1987. Data are reported at an hourly level of resolution, in which case one year's record for one site ideally contains 8760 values. For the purposes of this paper the hourly directional flow figures were aggregated into one directional flow per station per day; thus the basic time series comprised daily directional flow counts.

Gaps in the record posed a problem, since the ARIMA technique is intolerant of any missing values in the estimation and fitting process. Equipment malfunctions and other problems naturally produce a crop of missing values in the traffic counts. Data are particularly seriously censored in the winter months; December flow values are quite poorly represented. To test the ARIMA model we sought combinations of sites and years in which the traffic flow record was complete for at least three-quarters of the year with no more than 20 missing days of data. Then the missing values were supplied by interpolation. Since the variations in traffic flow are dominated by day-of-theweek effects, it seemed most natural to compute the missing values by interpolating flows from the previous and next days which were the same days of the week. Thus, a missing Tuesday value was replaced by the average flow of the closest two Tuesdays for which data were available. There did not appear to be a systematic basis in the types of sites (e.g., by traffic volume, or by urban/rural status) at which data were missing. From the acceptable sites, ten were chosen for analysis. Table 1 provides a brief description of each site; they are referred to here by their DOT reference number. Most were in the Capital District of New York where the snow study was concentrated, but sites were also chosen in the areas of Syracuse and Watertown.

GENERAL FLOW CHARACTERISTCS OF THE SITES

Broadly speaking, the stations analyzed here fall into two groups. The first type comprises sites with relatively little seasonal traffic variation and at which weekday travel usually predominates over weekend trips. It is reasonable to infer that flow at these sites

[2]J. Hayes and J. Pipkin, "The Disruptiveness of Snow and Ice Storms on Roadway Transportation in New York State. I A Snow Climatology and Research Design," AAG National Meetings, Baltimore, March 22, 1989; and J. Pipkin and J. Hayes, "The Disruptiveness of Snow and Ice Storms on Roadway Transportation in New York State. II Time Series Analysis of Traffic Flow Data," AAG National Meetings, March 22, 1989.

consists mainly of worktrips. Stations 1141, 7341, and 1446 are typical. Flow for station 1446 in one direction in 1980 is shown in the second half of Figure 1. This site is located at the eastern end of the principal bridge across the Hudson which carries Routes 9 and 20--and all westbound commuter traffic--into downtown Albany. Variants of this worktrip-dominated type are found at sites 7131 and 2431, which show a distinct summer peak as discretionary trips augment the usual flow of worktrips on weekdays, and come to dominate flow entirely at weekends.

The second type of site is dominated by discretionary travel, and typically shows strong weekend peaks. Stations 1552 and 1711 are typical. Flow at site 1711 for 1980 is shown in the lower half of Figure 2. This site is the first stop south of Albany on the NYS Thruway. There is more traffic at weekends than on weekdays, and there is a very pronounced summer travel peak, as well as noticeably higher travel at holiday periods such as "Presidents' weekend" in February (around Day 50) and Memorial Day weekend (around Day 150).

STRUCTURE OF ARIMA MODELS

Time series analysis has a long history in transportation science. Some important research foci are: models of speed and congestion relations;[3] predicting traffic volume in the short term (e.g. using Kalman filters)[4] and the long term (e.g. with a time-based logistic model; ARIMA/Box-Jenkins methods and spectral analysis)[5] and problems of missing values.[6] ARIMA (Box-Jenkins) modeling has not been widely applied in geography.[7]

The basic structure of ARIMA models is illustrated and described in Figure 1. This model may be motivated by a critique of OLS Regression as applied to timeseries data.[8] Discrete time series observations, say F_i, are assumed to be realizations of a stochastic process which includes error terms (or random shocks) satisfying statistical criteria identical to those of conventional regression analysis (i.e., the shocks are uncorrelated,

[3]e.g., F. Hall and T. Lam, "The Characteristics of Congested Flow on a Freeway Across Lanes, Space and Time," *Transportation Research* 22A (1988): 45-56; and D. Mahalel and A. Hakkert, "Time-series Model for Vehicle Speeds," *Transportation Research* 19B (1985): 217-225.

[4]I. Okutani and Stephanedes, "Dynamic Prediction of Traffic Volume Through Kalman Filtering Theory," *Transportation Research* 18B (1984): 1-11.

[5]S. Ahmed and A. Cook, "Analysis of Freeway Traffic Time-series Data Using Box-Jenkins Techniques," *Transportation Research Record* 733 (1979): 1-9; and H. Nicholson and C. Swann, "The Prediction of Traffic Flow Volumes Based on Spectral Analysis," *Transportation Research* 8 (1974): 533-538.

[6]Davis and N. Nihan, "Using Time-series Designs to Estimate Changes in Freeway Level of Service, Despite Missing Data," *Transportation Research* 18A (1984): 431-438.

[7]But see R. Bennet, "Process Identification of Time Series Modelling in Urban and Regional Planning," *Regional Studies* 8 (1974): 157-174; G. Clark, "Predicting the Regional Impact of Full Employment in Canada: A Box-Jenkins Approach," *Economic Geography* 55 (1979) 213-226; and L. Hepple, "Spatial and Temporal Analysis: Time Series Analysis," p0. 93-96 in N. Wrigley and R. Bennett (eds.), *Quantitative Geography: A British View* (London: Routledge and Kegan Paul, 1981).

[8]G.E.P. Box and G. Jenkins, *Time Series Analysis: Forecasting and Control* (San Francisco: Holden Day, 1976); D. McDowall, R. McCleary, E. Meidinger, and R. Hay, *Interrupted Time Series* (Beverly Hills: Sage, 1980); and R. McCleary and R. Hay, *Applied Time Series Analysis for the Social Sciences* (Beverly Hills: Sage, 1980).

homoscedastic, zero-mean, and normally distributed). The modeling approach represents each value of a time series as a function of three effects: autoregression, differencing and moving averages. These three effects are governed, respectively, by three structural parameters: *p, d,* and *q*, resulting in a model symbolized as *ARIMA(p,d,q).*

The *difference component* is used to model drift (as in a random walk) or trend (systematic increases or decreases in the F_i values with time). It is hypothesized that a given value of the series F_i equals the preceding value F_{i-1} plus a random step. If the mean of the random shock is zero, the result is termed drift; if it has a non-zero mean the result is trend. The process therefore integrates or accumulates over time. Such a process may be made homogeneous and stationary by differencing the values of the series one or more times. Thus, for example, $(F_i - F_{i-1})$ is the series of first order differences of Fi. The moving average component expresses a given F_i value as a weighted average of a specified number of previous observations. The persistence of past shocks is therefore finite; their effect eventually disappears from the system. The *autoregressive compoment* represents each observation as a linear function of a specified number of previous values. In particular the ARIMA (1,0,0) model represents the current observation as a portion of the immediately preceding one, plus a random shock.

These three components represent the *regular* (observation-to-observation) series. In addition a seasonal model may be specified in which the same three components operate at a specified periodicity. For our traffic data a seasonal model with a period of 7 was clearly mandated, since variations in the data are predominantly a day-ofthe-week effect (see Figures 1 and 2).

SOFTWEAR AND HARDWARE

Data were read from DOT-supplied tapes on a VAX and were downloaded to an AT (286) class PC. All analyses were performed using the time-series module TREND of the SPSS-PC program package, running with a math coprocessor.

MODEL FITTING PROCEDURE

Choice of an ARIMA model which is parsimonious yet which adequately fits data is not a mechanical or simple task. Ideally, ARIMA modeling proceeds through three stages: identification, estimation and diagnosis, which can be repeated as necessary. Various goodness of fit measures are available. In general the success of each modeling stage can be assessed by inspecting of the ACF (Autocorrelation Function) and the PACF (Partial Autocorrelation Function). The objective is to specify a model which reduces the residual ACF to white noise. As the fitting process proceeds, the ACF and PACF provide clues as to the type of model which is most appropriate. Indeed, standard texts on Box-Jenkins analysis contain normative ACF and PACF's for simulated data, to guide in the modeling choices.

At the outset, the time series must be stationary and homoscedastic; that is, means and variances must be constant. Stationarity can be obtained by various degress of differencing, while a logarithmic transformation often usually reduces heteroscedasticity. Almost all the traffic flow time series we examined were strongly nonstationary and heteroscedastic. Some experimentation indicated that a logarithmic transformation was always desirable. Differencing required more care. Over-differencing can do more harm than good (as reflected in irregular ACF and PACF plots). After considerable trial and error--which tended to confirm the notion that ARIMA modeling is an art which rewards trial and error and experience--the following method was procedure was applied.

First, the question of seasonality was addressed. That is, models of the form:

A(0,0,0)(sp,sd,sq) 7 ln.

were fitted. It is clear from applications reported in the literature that empirical *sp, sd,* and *sq* values almost never exceed 2. Seasonal difference values of 0, 1, and 2 were tested and the best value was chosen (i.e., the value that most reduces significant values in the ACF). Then the nine possible combinations of *sp* and *sq* (= 0, 1, and 2) were checked, and the best-fitting seasonal model was chosen. There are various different ways of assessing goodness fit in ARIMA. *SPSS Trends* output provides several statistics including two overall goodness of fit measures: the Akaike Information criterion and the Schwartz-Bayes criterion. An overall analysis of variance is available, and individual autoregressive and moving average coefficients have estimated t-values associated with them. Also each of the residual ACF series values has an associated standard error and a Box-Ljung statistic. Again after some experimentation, the "best" model was chosen as that which reduced the ACF series most nearly to white noise, as indicated by the individual standard errors and Box-Ljung statistics. The lowest 36 terms of the ACF were examined, and no more than two significant deviations from randomness were considered acceptable. In most case no more than one out of 36 was observed, which exceeds the 95% significance level. This procedure established the best seasonal (week-toweek) model, say (*sd*, sp*, sq**). Then, to find the best regular (day-today) model, the best seasonal model was combined with the "saturated" regular model:

A(2, 2, 2)(sd*,sp*,sq*) ln 7

and only those non-seasonal parameters were retained which differed significantly from zero according to the estimated t-values. This procedure was applied consistently to the test data, yielding the best-fit models presented in Table 3, where goodness of fit is measured by Mean Average Percent Error and Root Mean Square Error.

ARIMA replicated the traffic flow data very well. It compares favorably with polynomial regression (with day-of-the-week dummy variables), as reported else-where.[9] ARIMA greatly reduces the temporal autocorrelation, which is so clearly evident in the residuals of the polynomial regressions. We are currently exploring the power of ARIMA residuals to detect unusual traffic flow events, such as snowstorms.

EVALUATION

In the forecasting literature, ARIMA (Box-Jenkins) models are seen as relatively demanding of data, and relatively difficult to apply; that is, significant experience and/or trial and error is necessary to obtain the best fit. The ARIMA family of "complex and statistically sophisticated models"[10] generally perform well. However, they are by no means the best for all purposes. Makridakis and colleagues survey a bewildering variety of time-series methods each of which has strengths and weaknesses in: (1) analytical complexity (reflected in the need for computer time), (2) degree of expertise or judgment required to apply them, and, (3) fit to various kinds of data in short and long term

[9]Pipkin and Hayes, "Disruptiveness of Snow and Ice Storms, II."

[10]S. Makridakis et al. (eds.), *The Forecasting Accuracy of Major Time Series Methods* (New York: John Wiley, 1984).

forecasting. The work outlined here justifies several conclusions about ARIMA in traffic flow applications.

First, ARIMA models provide excellent descriptions of daily traffic flows on part of the NYS highway system. Judged by the visual appearance of the fit (the Figures show two typical examples), as well as by a formal array of goodness-of-fit statistics, ARIMA provides fits which are excellent in absolute terms, and which are relatively better than polynomial regressions.

Second, the flow series are typically non-stationary, but first order differencing, both regular and seasonal, was sufficient in almost all cases to produce stationarity. In only one case, site 1446, a major river bridge into downtown Albany, was second order seasonal differencing necessary. In general, seasonal differencing was more useful than regular differencing. This is especially true at those sites which exhibit a summer peak in discretionary and weekend travel, so that the annual profile of traffic low approximates the profile of a positive quadratic function.

Third, to examine various model specifications
exhaustively is not prohibitively time-consuming, even using a PC statistics package. Two strategies were tested here: (1) separate examination of each possible autoregressive and moving average model, for the stationary series, after any required differencing has been done, and (2) specification of a "fully saturated" model (after differencing) with retention of only those autoregressive and moving average parameters that are significantly different from zero. The fundamental reason why these strategies are feasible is that the orders of autoregression and moving averages, as well as of differencing, can be assumed not to exceed 2. Thus there is no combinatorial explosion of possibilities that must be evaluated individually.

Fourth, an interesting question arises regarding the substantive interpretation of autoregressive and moving average effects. Are the different flow regimes of worktrip- and nonworktrip-dominated highways reflected in the predominance of either autoregressive or moving average effects? In representing the seasonal (lag 7) contributions to flow, the majority of sites record significant moving average rather than autoregressive coefficients. Thus, flows seem to be more dependent on previous shocks or random effects than on previous flows. With the single exception of site 1711, the sites which record significant seasonal autoregressive effects are work-trip dominated. In these cases previous flows, rather than previous shocks, are the best predictors of new flows. At the level of regular (non-seasonal flows) both autoregressive and moving average effects tend to be significant. Finally, it should be pointed out that this exploratory analysis barely scratches the surface of geographic research that could be done on the extremely rich DOT traffic flow data. What kinds of classifications of flow regimes can be developed? How do flow patterns reflect the geography of local network structure? Do the general conclusions on ARIMA modeling obtained here apply when flow data are analyzed at an hourly rather than a daily level?

A particularly interesting application of ARIMA modeling is in impact assessment. In New York State --and particularly in places such as the Albany Capital District, where traffic congestion is burgeoning both in fact and in public perception--traffic flow regimes are likely to be a resultant of (1) steady, secular growth which can be described by a trend or differencing model, and (2) one-time "shocks" which permanently change travel patterns, including policy actions prompted by congestion or by new residential or commercial developments. Clearly a new access ramp or a new road will affect traffic patterns in measurable ways. But what about relatively small changes, for example in signage or

routing? *Interrupted time series analysis*[11] applies ARIMA and other models to detect such effects. This is one of the many possibilities for future research on large traffic flow data sets.

TABLE 1

Structure of ARIMA Models

Regular Models

General Time (Flow) Series F_1, F_2, F_3, ... incorporating error terms or random shocks e_1, e_2, e_3, ... with:

$E(e_i) = 0$
$VAR\ (e_i) = s^2$
$COV(e_i,e_j) = 0,\ i<>j$
e_i is Normal

[11] D, McDowall et al., *Interrupted Time Series*.

TABLE 1 (cont.)

The general ARIMA model is represented as

A(p,d,q) where:
p = autoregressive order
d = order of differencing
q = order of moving average

Differencing

A(0,1,0) $F_i - F_i\text{-}1 = e_i$, or $F_i = F_{i-1} + e_i$

A(0,d,0) involves differencing *d* times

Autoregressive Process

A(1,0,0) $F_i = a_1 F_{i-1} + e_i$

A(2,0,0) $F_i = a_1 F_{i-1} + a_2 F_{i-2} + e_i$,

with autoregressive parameters a_1, a_2.

Moving Average

A(0,0,1) $F_i = e_i - b_1 e_{i-1}$

A(0,0,2) $F_i = e_i - b_1 e_{i-1} - b_2 e_{i-2}$,

with moving average parameters b_1, b_2.

Seasonal Models

In addition to the regular components above, a seasonal model may be applied to capture any predictable fluctuations with a known period, with seasonal orders of differencing, autoregression and moving averages, sd, sp, and sq. The most general model with seasonal lag n is written as:

A(p,d,q)(sp,sd,sq) n.

For example, the pure seasonal, first-order autoregressive model A(0,0,0)(1,0,0) 7, (which represents a weekly seasonality for daily observations), states that:

$F_i = a_1 F_{i-7} + e_i$.

Two other structural choices that must be made are whether the regular and seasonal components combine additively or multiplicatively, and whether the model bests fits intransformed or logarithmically transformed data. The general multiplicative, logarithmic model used throughout thus study is written:

A(p,d,q)(sp,sd,sq) ln 7.

TABLE 2

Description of Sites

(4-digit values are DOT station reference numbers.)

1141 on Rt. 9, .1 miles South of Rt. 155, Principal Arterial, Urban, 4 lanes

1711 on I-87, 1.7 miles North of Exit 22, Access of Rts. 9 and 9N, Interstate, Rural, 5 lanes

1351 on Rt. 9W, .1 miles South of Rt. 81, Minor Arterial, rural, 2 lanes

1552 on Rt. 147, 2 miles South of Rt. 67, Major Collector, Rural, 2 lanes

1446 East End of Dunn Memorial Bridge, Principal Arterial (Expwy.), Urban, 5 lanes

1470 on CR 40, .3 miles East of East End of Rt. 134, Major Collector, Rural, 1 lane

2431 on Rt. 31, 1.2 miles East of the Onondaga-Madison County line, Minor Arterial, Rural, 2 lanes

7341 on Rt. 11, 3 miles South of Rt. 3, Watertown, Minor Arterial, Urban, 4 lanes

Source: *Regional Traffic Volume Data from Continuous Count Stations in New York State: 1972-1982*, Transportation Statistics and Analysis Section, Data Services Bureau, New York State Department of Transportation, Albany, NY 12232, July 1983.

TABLE 3

Best-Fit ARIMA Models

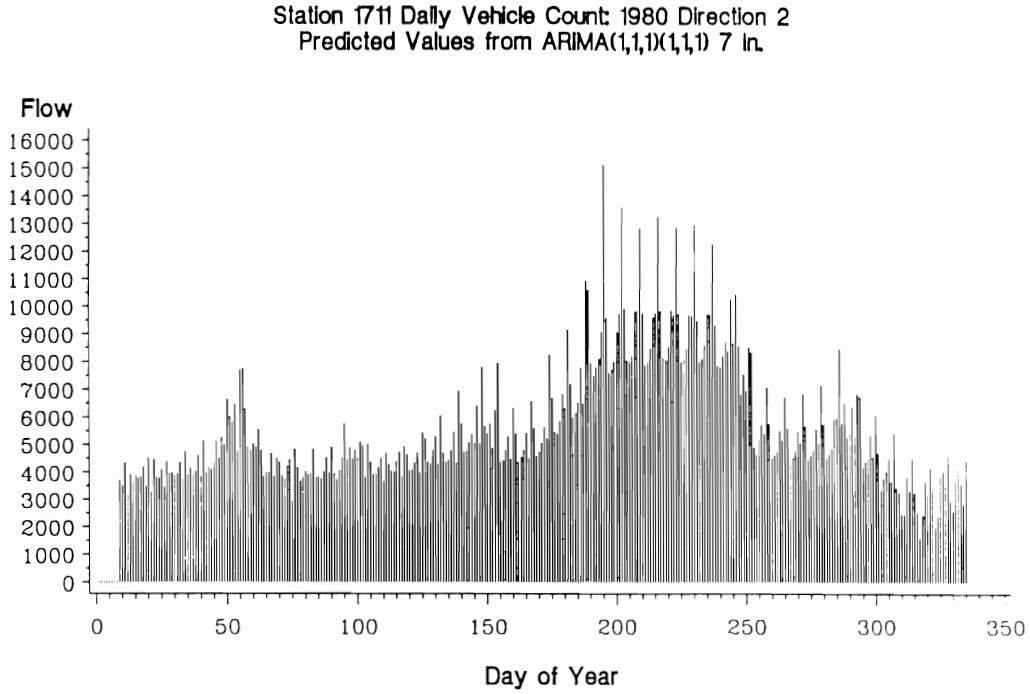
In each case the dependent variable is the logarithm of flow. ARIMA models have regular and seasonal components (n = 7). *Site* refers to the DOT count station identifier. *Year* indicates the date of the flow data. *Dir* indicates the direction traffic flow.

Site	Year	Dir	ARIMA MAPE	ARIMA RMSE
			Best Model (1,1,1) (0,1,1)	
2431	87	1	.0014	.7160
1552	87	1	.0068	1.4222

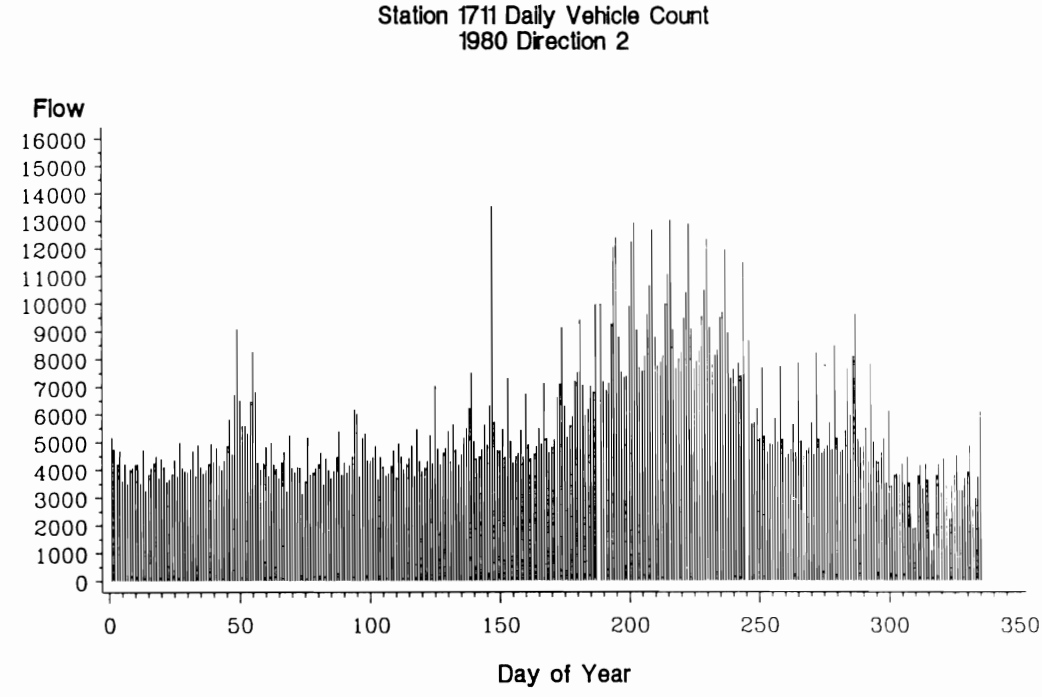
TABLE 3 (cont.)

1141	86	2	.0005	.8236
1552	80	1	.0068	1.3197
1351	87	1	.0024	.8782
1552	87	2	.0059	1.3795
1552	80	2	.0065	1.3533
			Best Model (0,1,1) (0,1,1)	
1470	80	1	.0034	.6718
1351	86	1	.0022	.7984
			Best Model (1,0,1) (0,1,1)	
7341	86	1	.0011	.6256
1351	86	2	.0019	.5920
1446	80	1	.0006	.6267
1141	86	1	.0007	.6091
1470	80	2	.0032	.4748
			Best Model (0,1,1) (0,1,1)	
1711	80	2	.0023	1.9881
1351	87	2	.0021	.0765
1351	86	2	.0019	.5920
			Best Model (0,1,1) (0,1,1)	
1446	80	2	.0006	.4748

MAPE = Mean Absolute Percent Error
RMSE = Root Mean Square Error



Data Courtesy of New York State
Department of Transportation



Data Courtesy of New York State
Department of Transportation