Research Report No. UVACTS-13-0-78
May, 2003

Center for Transportation Studies
at the University of Virginia

Imputation Techniques to Account for Missing Data in Support of Intelligent Transportation Systems Applications

By:

Linh N. Nguyen
Dr. William T. Scherer

A Research Project Report for the Virginia Department of Transportation (VDOT)

Linh N. Nguyen
University of Virginia

Dr. William T. Scherer
Department of Systems and Information Engineering
University of Virginia
Email: wts@virginia.edu

Center for Transportation Studies at the University of Virginia produces outstanding transportation professionals, innovative research results and provides important public service. The Center for Transportation Studies is committed to academic excellence, multi-disciplinary research and to developing state-of-the-art facilities. Through a partnership with the Virginia Department of Transportation's (VDOT) Research Council (VTRC), CTS faculty hold joint appointments, VTRC research scientists teach specialized courses, and graduate student work is supported through a Graduate Research Assistantship Program. CTS receives substantial financial support from two federal University Transportation Center Grants: the Mid-Atlantic Universities Transportation Center (MAUTC), and through the National ITS Implementation Research Center (ITS Center). Other related research activities of the faculty include funding through FHWA, NSF, US Department of Transportation, VDOT, other governmental agencies and private companies.

CTS Website
http://cts.virginia.edu

Center for Transportation Studies
University of Virginia
351 McCormick Road, P.O. Box 400742
Charlottesville, VA 22904-4742
434.924.6362

1. Report No.

UVACTS-13-0-78

2. Government Accession No.

3. Recipient's Catalog No.

4. Title and Subtitle
Imputation Techniques to Account for Missing Data in Support of Intelligent Transportation Systems Applications

5. Report Date
May, 2003

6. Performing Organization Code

7. Author(s)
Linh N. Nguyen, William T. Scherer

8. Performing Organization Report No.

9. Performing Organization and Address

Center for Transportation Studies
University of Virginia
PO Box 400742
Charlottesville, VA 22904-7472

10. Work Unit No. (TRAIS)

11. Contract or Grant No.

12. Sponsoring Agencies' Name and Address
Office of University Programs, Research and Special Programs Administration
US Department of Transportation
400 Seventh Street, SW
Washington DC 20590-0001

13. Type of Report and Period Covered
Final Report

14. Sponsoring Agency Code

15. Supplementary Notes

16. Abstract

This research will address the feasibility and applicability of imputing missing traffic data, and perform an analysis on five imputation techniques. The first technique examined in this study is historical average imputation. The second two techniques are new heuristic approaches designed specifically for missing traffic data. Finally, the last two are statistical imputation procedures, which have been developed and implemented in various other fields. These approaches for imputing missing data are implemented and compared in this study.

17 Key Words
imputation, missing data, databases, statistical techniques

ABSTRACT

Applications of Intelligent Transportation Systems (ITS) seek to improve efficiencies in transportation by using emerging information technologies. One element of the ITS infrastructure is traffic signal control used to improve traffic flow over arterial road networks. Traffic signal control systems take advantage of a variety of advanced control strategies that take into account current traffic conditions.

The most widely used form of traffic surveillance device is the single-loop detector. System detectors are single-loop detectors that are placed well behind stop bars at most signalized (arterial) intersections, and used to capture traffic conditions.

There are several important applications of detector data. Traffic engineers develop signal-timing plans using historical data to control over 900 signalized intersections, in the case of the Virginia Department of Transportation's (VDOT) Smart Traffic Signal Systems (STSS) group. However, a more urgent requirement for detector data exists for automated traffic signal control systems, such as the traffic-responsive (TRSP) mode in first-generation control (1-GC) systems. These systems control and implement suitable signal-timing plans for arterial networks based on current traffic conditions that are described by detector data.

There is an inherent reliability problem with surveillance devices such as single-loop system detectors. These types of detectors are prone to fail and can be attributed to many natural and man-made factors. The situation of detectors going off-line would render useless automated signal control systems.

The objective of this research is to investigate imputation techniques for estimation of missing values of time-critical traffic data due to off-line or non-responsive system detectors. Ideally, imputation techniques should exploit the underlying spatial relationships among system detectors throughout an arterial network. Estimates of missing system detector data at any particular moment should be representative of the arterial network's traffic conditions to effectively support continuous operations of traffic-responsive or traffic-adaptive signal control strategies.

This research proposes a new imputation model class called the C-STARMA. This new development is based upon the Space-Time Series model (STARMA) but extends that technique to incorporate contemporaneous data in addition with historical time series data. The C-STARMA model building procedure is also described and implemented using real-world traffic data. The C-STARMA was empirically validated to consistently outperform other techniques implemented in this research. We observed that this new model class performed well at estimating traffic data (volume and occupancy), as well as showing a high level of precision.

ACKNOWLEDGEMENTS

I am fortunate to have returned to the Department of Systems Engineering, from which I was trained as an undergraduate engineer, to pursue advanced studies. I extend my deepest appreciation and gratitude to Professor William T. Scherer for accepting me as a graduate student and his research assistant. Professor Scherer always provided sound guidance, and I thoroughly enjoyed the intellectual challenge coming forth from our many exploratory discussions.

I am indebted to Professor Brian L. Smith for inviting me to the Smart Travel Lab research team; to Professor Phillip E. Pfeifer for sharing his insight on the STARMA methodology; and to Professor James W. Lark for serving as my Committee Chairman.

Many individuals have made my research possible, and I thank these people: James Conklin for his thorough draft reviews and being my tireless connection back in C-ville; Keith Oswald for acting as a sounding board to voice my ideas; Dave Lewis and Xiaoning Lu for impeccable STL operations; the S.E. Department Staff (Sharon Predmore, Dee McCauley, Margaret Proffit) for always lending their valuable assistance.

Finally, I extend kindest and sincerest thanks to my parents and family, Trinh T. Tang and Lan N. Nguyen; to my fiancée Kathryn Aitkenhead; and to my colleagues and friends for their love, encouragement, and support throughout my tenure as a graduate student.

L. Nguyen
May 2003

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1.0 INTRODUCTION

1.1 Intelligent Transportation Systems

The development of ITS systems seeks to improve efficiencies in transportation by using emerging information technologies. One element of the ITS infrastructure is traffic signal control to improve traffic flow over arterial road networks. Traffic signal control systems take advantage of a variety of advanced control strategies that take into account current traffic conditions.

1.1.1 Advanced Traffic Signal Control Strategies

The most widely used form of traffic surveillance device is the single-loop detector. System detectors are single-loop detectors that are placed well behind stop bars at most signalized (arterial) intersections, and used to observe current traffic conditions. Data is collected at 1-minute intervals, however, it is aggregated and archived to the database every 15-minutes. The database stores data for volume (vehicles per hour), occupancy (percent of hour), and speed (miles per hour) according to time-of-day and date.

There are several important applications of detector data. Traffic engineers develop signal-timing plans using historical data to control over 900 signalized intersections in the case of the Virginia Department of Transportation's (VDOT) Smart Traffic Signal Systems (STSS) group. However, a more urgent requirement for detector data exists for automated traffic signal control systems, such as the traffic-responsive (TRSP) mode in first-generation control (1-GC) systems. These systems control and implement suitable signal-timing plans for arterial networks based on current traffic conditions that are described by detector data. Data on current traffic conditions can be captured by other surveillance technologies as well, including video cameras and ramp meters.

There is an inherent reliability problem with surveillance devices such as single-loop system detectors. At any given time, approximately 25-30% of the detectors are off-line. Detector failures can be attributed to many natural and man-made factors [32]. For instance, wires connecting detectors to control boxes may corrode due to water penetration or may be severed by construction teams digging into the ground. This reliability problem greatly affects the signal control systems that require timely and accurate detector data in order to properly function.

1.2 Rationale

Traffic operations centers, such as the VDOT STSS, use data collected from system detectors to support critical traffic management operations. A continuous feed of system detector data capturing current traffic conditions enable advanced signal control systems to function properly. Due to these systems' heavy reliance on detection devices, VDOT management has foregone the use of signal control strategies that may be more efficient than the ones employed in current practices.

Traffic management systems, both currently employed and next generation systems that rely upon traffic detection, do not include mechanisms to ensure effective continuous operation in the event of surveillance system failure or non-response. In addition, current research does not specifically address how advanced signal control systems should operate in event of failure or non-response.

1.3 Goals and Objectives

The objective of this research is to investigate imputation techniques for estimation of missing values of time-critical traffic data due to off-line or non-responsive system detectors. The objective is to support continuous operations of advanced signal control systems as a fault-tolerant system. Estimates of missing system detector data should be representative of current traffic conditions to effectively support continuous operations of traffic-responsive or traffic-adaptive signal control strategies.

1.4 Scope

The scope of this study is limited to estimating values of vehicular volume (vehicles per hour) and occupancy (percent) within Northern Virginia's arterial networks. These values are normally observed using traffic surveillance devices, such as system detectors embedded along approaches to signalized intersections.

This study does not examine the characteristics or patterns of missing data from system detectors. That is, we do not seek to characterize the reliability of surveillance devices, such as single-loop detectors. Our purpose is to estimate missing values of traffic data captured by these detection devices in the event of their failure.

The study focuses upon a sub-network of intersections within the greater Reston Area Network (RAN) in Fairfax County, Virginia. This sub-network comprises of the three adjacent intersections of Reston Parkway and South Lakes Drive, Glade Drive, and Fox Mill Road.

1.5 Overview of Technical Report

This thesis is organized in seven chapters, including the introductory chapter. The remaining chapters are summarized as follows:

- *Chapter 1: Introduction*

- *Chapter 2: ITS Advanced Signal Control Systems*

This section discusses in detail the current state of practice in the domain of traffic signal control systems. Different signal control strategies and the information systems implementing these strategies are presented. In particular, their heavy reliance on current traffic data that is captured by detection devices is highlighted.

- *Chapter 3: Data Imputation Techniques*

A brief discussion is presented on traditional computation techniques to treat missing data. These techniques are primarily of interest to the analyst who wishes to analyze a data set that contains missing data, by substituting imputed values for missing values. Since this research project's scope is imputing data in real-time, data forecasting and estimation techniques are presented as the computational basis for the study.

- *Chapter 4: Problem Formulation & Implementation*

This section presents the mathematical formulation of the data imputation techniques discussed in chapter 3. Also, the implementation details of these techniques are presented.

- *Chapter 5: C-STARMA*

This section presents the new model class by describing its development, notation, and model building procedure.

- *Chapter 6: Results & Analysis*

The results of the models' effectiveness to estimate current traffic data are presented and analyzed. The proposed techniques are evaluated in terms of their viability for operational use. Issues to consider include computational complexity and performance, ease of coding, and other practical aspects.

- *Chapter 7: Conclusion*

The final chapter summarizes the research findings and contributions, and presents recommendations for further research.

2.0 ITS ADVANCED SIGNAL CONTROL SYSTEMS

2.1 Data Requirements in Intelligent Transportation Systems

Traffic operations centers, such as VDOT's Northern Virginia Traffic Signal Systems center, use data collected from system detectors to support critical traffic management operations. In VDOT's case, system detector data are collected via single-loop detectors placed well behind stop bars at most signalized (arterial) intersections. This data describes the traffic conditions at 15-minute intervals and are used by traffic engineers to develop timing plans for signal control of its 900+ signalized intersections. Thus, detector data is a necessary ingredient in developing appropriate timing plans for first-generation traffic signal control systems. Next-generation systems such 2-GC and 3-GC that implement real-time adaptive signal control rely even more on system detectors to supply data on current traffic conditions.

2.1.1 Traffic Signal Control Systems

Traffic engineers adopted the use of computer-based traffic signal control systems in the 1960's to develop more responsive signal control strategies. One of the most comprehensive studies of new signal control strategies was the Urban Traffic Control System (UTCS) conducted by the Federal Highway Administration (FHWA) in the 1970's (Gartner et al. 1991). The purpose of the UTCS project was to develop and test a variety of advanced network control concepts and strategies. This project spanned over almost a decade and its results defined the state of the art in the United States until present-day. Research and testing of signal control strategies in the UTCS project were divided into three generations: first-generation control, second-generation control, and third-generation control.

2.1.1.1 First Generation Control (1-GC)

The widely implemented type of signal control strategy is First-Generation Control (1-GC). This type of system uses pre-determined signal timing plans that are developed off-line based on historical traffic data, such as volume (vehicle per hour). Traffic management centers implementing 1-GC systems are given several options to select timing plans to control the network: time-of-day (TOD), manual, and traffic-responsive (TRSP). Example systems include UTCS, Series 2000, and MIST.

2.1.1.2 TOD Mode

Traffic engineers develop signal timing plans for assumed traffic patterns based on historical data. Signal plans are implemented by programming the signal control hardware to execute specific timing plans at designated time intervals. For example, an AM signal plan can be designed to handle morning traffic between 0600 and 1000 o'clock.

2.1.1.3 MANUAL Mode

Traffic management centers can manually select a pre-defined signal timing plan from its stored library to handle special traffic conditions. For example, a special timing plan can be implemented in lieu of the TOD plan if traffic conditions at a particular location and moment exceeds the normal traffic patterns.

2.1.1.4 TRSP Mode

Traffic-responsive automatically selects and implements the signal timing plan which is best suited to current traffic conditions. A number of timing plans for various traffic conditions are developed off-line and stored in the timing plan database. Traffic surveillance devices, such as loop detectors, measure the volume and occupancy of current traffic conditions. The signal controller then implements the plan from the database that has the characteristics that best matches current conditions. TRSP mode updates signal plan selection in 15-minute cycles.

2.1.1.5 1.5 Generation Control

1.5-GC systems fill the gap between 1-GC and 2-GC systems, in that timing plans are developed on-line according to current traffic conditions. These timing plans are not automatically implemented, however they are stored in the plan database for the traffic engineer to select.

2.1.1.6 Second Generation Control (2-GC)

Second-Generation Control is an on-line strategy that computes in real-time and implements signal timing plans based on traffic surveillance data and predicted values. This optimization process can be repeated at 5-minute intervals. However, to avoid too many transitions between plans, new timing plans cannot be implemented more often than once every 10 minutes. Example systems are SCOOT and SCATS.

2.1.1.7 Third-Generation Control (3-GC)

Third-Generation Control strategy was designed to realize a fully responsive, adaptive, on-line traffic control system. The difference from 2-GC was that the period after which timing plans were revised was reduced to 3 to 5 minutes. Example implementations include RT-TRACS and RHODES.

2.1.2 Freeway Systems Engineering

The widespread adoption of advances in intelligent transportation systems enables traffic management centers to provide useful information to motorists in their respective jurisdictions. Motorists can be informed by such means as variable message signs (VMS) of downstream incidents, congestion, expected travel-times along major arteries, and alternative routes to avoid congestion.

2.1.2.1 Incident detection & management

Incident management is the coordinated, preplanned use of human and technological resources to restore full capacity of arterial or freeway roadways after an incident occurs. This also includes providing timely and relevant information and direction to motorists until the incident is cleared. In order to be effective, incident management programs must reduce the following: time to detect an incident, time to identify the nature of an incident, time to respond and clear the incident, and traffic demands during the incident by applying tactical traffic management measures (e.g. re-routing traffic). Rapid detection is a key element in incident management in determining that an incident has occurred and to minimize its impact on roadway capacity.

One technique for incident detection uses electronic surveillance, such as loop detectors. These sensors are placed along the roadway at predetermined intervals and detect the presence of vehicles. This data is processed automatically to determine roadway congestion.

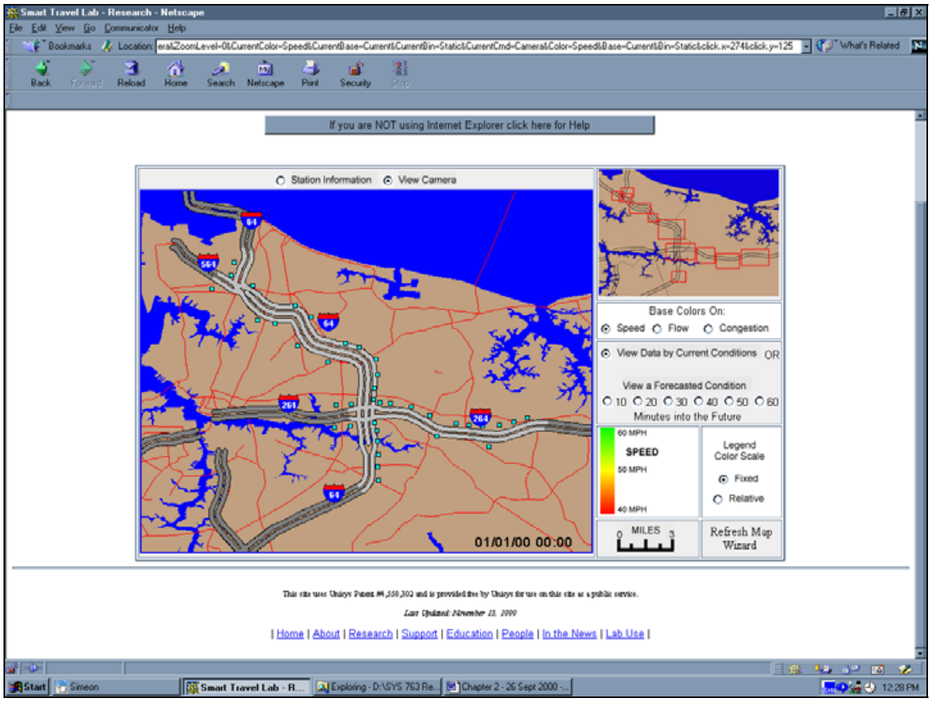
The advantages of using this type of detection include the ability to continuously monitor entire roadway sections and to provide rapid detection, especially in high-volume conditions. However, the trade-offs are high costs associated with planning, design, installation, operations, and maintenance of the detectors. Electronic surveillance also does not perform well at detecting non-congestion-causing incidents.

2.1.2.2 Travel-time estimation

Several traffic management centers in the U.S. have implemented mechanisms to inform motorists of expected travel-times along major arteries. The algorithms that estimate travel times base their calculations upon traffic surveillance data recorded by such devices as loop detectors embedded in the roadway. This information is disseminated to motorists typically via variable message signs along the freeway. An example implementation is the Georgia DOT and Seattle area freeway.

2.1.2.3 Congestion maps

Most ITS professionals would agree that the effectiveness of incident detection has been rather poor. Laboratory research is investigating alternative approaches to automated condition monitoring. The research is departing from the traditional goal of detecting incidents to attempting to detect if the system is operating "out-of-normal" range regardless of an incident state (http://smarttravel.virginia.edu).



If you are NOT using Internet Explorer click here for Help

This site uses Unisys Patent #4,558,302 and is provided free by Unisys for use on this site as a public service.

Last Updated: November 15, 1999

Figure 2-1: Congestion map for Hampton Roads (Source: http://smarttravellab.virginia.edu)

2.2 Missing Data in the Traffic Engineering Domain

Missing data in the traffic engineering domain is a critical issue that has yet to be definitively investigated. Traffic surveillance data, such as loop detector data, enable traffic management centers to monitor current traffic condition in their jurisdiction. Loop detectors provide data on vehicle count (volume per hour), occupancy (percentage of hour), and estimated average speed along the road link. This data is essential to traffic engineers tasked with developing signal timing control strategies. Real-time data is also critical to the operations of UTCS systems such as traffic-responsive mode or adaptive-control systems.

Reliability is a primary concern with surveillance devices such as single-loop system detectors. At any given time, approximately 25-30% of the detectors are off-line and contribute to missing data problems in the traffic engineering domain. Detector failures can be attributed to many natural and man-made factors (Parsonson 1984) (Patel 1995). For instance, wires connecting detectors to control boxes may corrode due to water penetration or may be severed by construction teams digging into the ground. This reliability problem greatly affects the signal control systems that require timely and accurate detector data in order to properly function.

System detector data is archived in the Smart Travel Lab's database in 15-minute intervals throughout the day, spanning from 0:00 to 23:45 hours. Detector data includes volume (vehicles per hour), occupancy (percent of hour), and speed[1] (mph) values associated with individual system detectors. There are approximately 900+ system detectors for which the database archives historical data. Data collection for the northern Virginia (Fairfax County) region was initiated in February 2, 2000.

Table 2-1: Sample System Detector Data

Date / Time	2026 Vol	2026 Occ	2027 Vol	2027 Occ	2028 Vol	2028 Occ
4/3/00 9:00	124	2	276	2	380	2
4/3/00 9:15	120	8	284	2	252	1
4/3/00 9:30	124	2	256	2	300	2
4/3/00 9:45	112	1	244	2	256	2
4/3/00 10:00	140	2	252	2	224	2
4/3/00 10:15	120	1	320	3	364	3
4/3/00 10:30	92	1	300	3	212	1
4/3/00 10:45	108	1	332	3	292	2
4/3/00 11:00	132	2	332	3	252	2
4/3/00 11:15	136	2	312	3	296	2
4/3/00 11:30	164	2	396	3	336	3
4/3/00 11:45	124	2	420	4	380	3
4/3/00 12:00	204	5	496	5	448	4

---

[1] Speed, or Average Speed (mph) is a calculated value based upon volume and occupancy data. This research focuses only upon estimating missing values of volume and occupancy.

The occurrences of missing data in the traffic management domain can be described using several scenarios. Missing data may occur for short time periods, on the order of one to several 15-minute time intervals. This may be caused by a slight malfunction in the detector hardware, software, or communications line between the detector, controller station, and central computer system at the traffic management center. This particular scenario exhibits detector data that is missing for several time intervals but data availability is resumed within a short time period (less than one hour).

Longer intervals of missing data, upwards of several hours to days' worth of data, may be attributed to failure in the detector, computer system, or communications components. In such cases, response maintenance crews are deployed to resolve technical difficulties and to get the detectors back on-line. System failures occurs due to man-made factors, such as communications lines being severed by construction, or due to natural factors like electrical power surges or short-circuits due to moisture in the loop detector hardware. Depending upon the availability of the maintenance crew or urgency to fix the detection system, detector data may not resume for a longer period of time (on the order of days). In addition, traffic management centers or maintenance crews may perform routine operations and maintenance activities on detection systems such as restarting hardware systems or hardware/software upgrades and, thereby, necessitating taking detectors off-line.

2.2.1 Temporal attribute of detection data

Detection data in advance traffic management systems feature both temporal and spatial attributes. In terms of its temporal nature, detection data (speed, volume, occupancy) are collected on a 1-minute interval and are aggregated to the database in 15-minute intervals based upon time-of-day interval and date. Occurences of missing data include unavailability of traffic data for a single TOD interval, short spans (several TOD intervals), or longer durations (more than one day). Missing data can occur for a single detector or multiple detectors within the network.

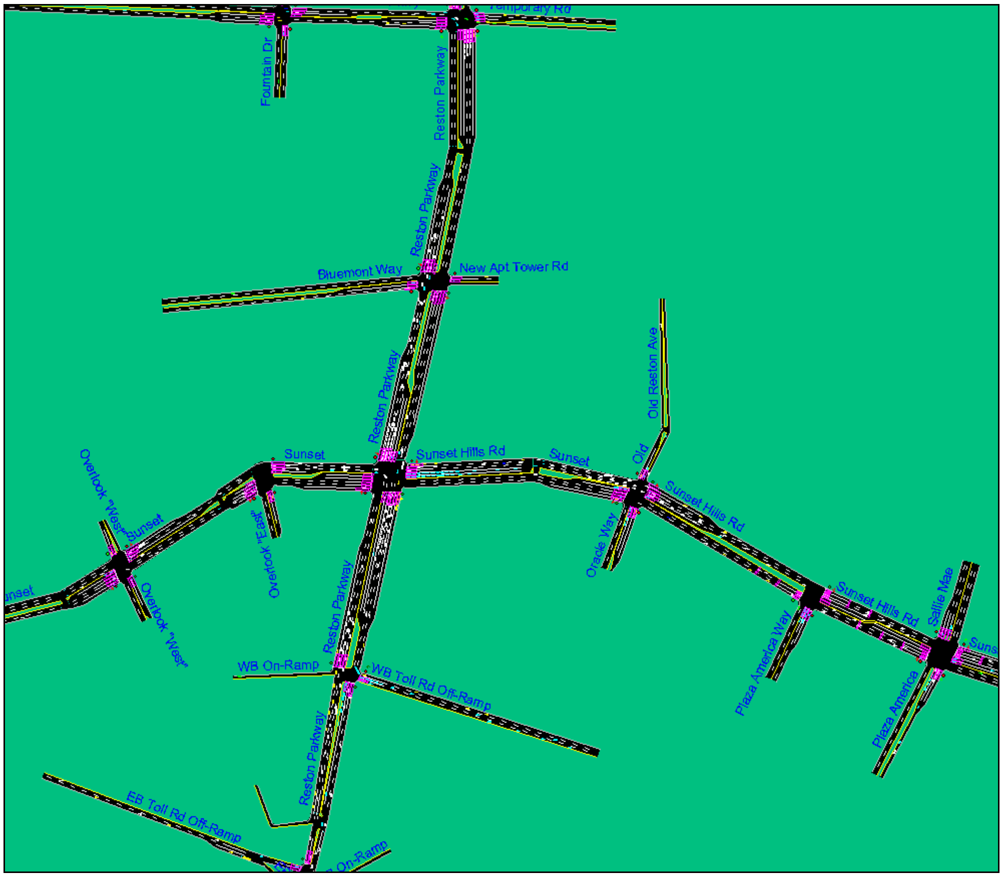


Figure 2-2: Partial View of Reston Area Network (Source: SimTraffic simulation by VDOT STSS)

2.2.2 Spatial attribute of detection data

Detector data also exhibit spatial characteristics due to their geographic placement throughout the network. A network consists of arterial road intersections (includes main throughway and sidestreets) for which traffic engineers develop signal timing plans. System detectors are situated along each approach to the intersections within our sub-network of the Reston Area Network. These detectors capture lane-specific data for traffic movements along the corridor and side streets. Thus, data collected at any intersection is expected to exhibit spatial correlation with the data collected at other intersections in the same network due to their proximity.

Our research examines the situation when traffic data is missing for a single detector, and then estimate missing data for this one location. We shall investigate three particular network scenarios that use available detectors as model inputs: upstream detectors only, both upstream and downstream, and downstream detectors only.

3.0 DATA IMPUTATION TECHNIQUES

3.1 Statistical Methods On Treating Missing Data

There has been significant research and literature on how to handle missing values from data sets. Techniques to account for missing data range from simple heuristics to complex data estimation methods, such as the following:

- *Mean of overall series*: This technique simply substitutes the statistical mean of observed values in the data set for the instances of missing values.
- *Mean of period within the series in observation is missing*: The mean of observed values within a specified period is substituted for missing values that occur within that period.
- *Mean of adjacent observations*: This algorithm allows the analyst to specify the "sliding window" size and computes the statistical mean of using observations before and after the interval of missing values.
- *Interpolation*: These algorithms replace missing values by interpolating from previously observed values. Such techniques include moving averages, exponential smoothing, linear splines, cubic splines, etc.
- *Regression Imputation*: These models fill in the missing values using predicted values from a regression of a given variable on other variables in the analysis.

- *Time Series*: If the data set is comprised of observations at a given interval of time, then these forecasting techniques are powerful at estimating values for one to several intervals into the future.

Missing data values may be commonplace in data collection efforts, such as social surveys or scientific experiments, as well as in intelligent transportation system data archives. This can be attributed to numerous factors, which include non-response from the study's sample or malfunction of data collection devices. Classical approaches to estimating missing data values, such as the ones above, are sufficient for studying incomplete data sets when the analyst wishes to account for these instances. Gold et al. (2001) proposed several methods for imputing non-response in traffic volumes occurring in intervals under five minutes. They applied imputation methods ("factor up" and straight-line interpolation) and two regression methods (polynomial and kernel) to estimate missing values of traffic volume within the ITS database. These techniques employed available data prior to and subsequent of intervals of missing values. For example, Table 3-1 shows sample traffic data collected by detectors 2001, 2002, and 2002. The Gold et al. study focused on filling in the missing data that occurred for detector number 2020 using previous and subsequent values to the empty intervals.

Table 3-1: Instances of Missing Data within an ITS Database

DetectorID	2001			2002			2020		
DateX	**Volume**	**Occupancy**	**Speed**	**Volume**	**Occupancy**	**Speed**	**Volume**	**Occupancy**	**Speed**
6/19/00 16:00	492	5	44	364	3	43	340	12	24
6/19/00 16:15	512	5	45	352	3	47	328	13	23
6/19/00 16:30	500	11	37	396	5	42	396	20	16
6/19/00 16:45	496	5	44	372	3	48	344	14	20
6/19/00 17:00	424	4	45	324	3	52			
6/19/00 17:15	504	5	44	412	4	43	420	24	16
6/19/00 17:30	464	6	38	372	3	47			
6/19/00 17:45	544	6	44	436	4	48			
6/19/00 18:00	532	7	39	408	4	48	380	11	22
6/19/00 18:15	452	4	43	368	3	45	324	17	21
6/19/00 18:30	396	4	44	340	3	47	244	4	29

However, the focus of this thesis project is to estimate in real-time missing data values from non-responding or malfunctioning system detectors that directly feed into traffic-responsive or traffic-adaptive signal control systems. Thus, the methods proposed in this research only take advantage of traffic data leading up to the interval(s) of non-response. This estimation is to be performed in real-time to support on-line signal control systems and, therefore, can only apply available data up to the moment of the missing observation. Table 3-2 illustrates this concept: we need to estimate the data for the latest interval for which non-response occurs for detector 2020.

Table 3-2: Example of Estimating Missing Values for the Current Time Interval

DetectorID	2001			2002			2020		
DateX	**Volume**	**Occupancy**	**Speed**	**Volume**	**Occupancy**	**Speed**	**Volume**	**Occupancy**	**Speed**
6/20/00 15:00	440	5	38	380	3	43	192	2	36
6/20/00 15:15	420	4	41	360	3	49	208	3	38
6/20/00 15:30	408	4	46	348	3	50	196	4	31
6/20/00 15:45	440	4	48	400	3	50	232	3	34
6/20/00 16:00	408	4	43	316	2	50	244	3	29
6/20/00 16:15	424	4	44	372	3	52	288	5	27
6/20/00 16:30	388	4	49	392	3	51	272	5	28
6/20/00 16:45	456	4	45	356	3	53	324	15	25
6/20/00 17:00	460	4	45	368	3	50	444	19	20
6/20/00 17:15	436	4	46	432	4	45	364	21	20
6/20/00 17:30	484	7	39	424	5	41			

3.2 Data Estimation Techniques

Classical approaches can perform very well at estimating missing values for incomplete data sets. We wish to apply several of these techniques to estimate missing values in the ITS domain of real-time signal control systems. Again, the scope of this research is to investigate the applicability of several imputation techniques to estimate for missing values. We did not perform an exhaustive evaluation of the plethora of data imputation methods. However, the following techniques were selected due to the intuition that traffic data in the ITS domain is quite suitable to be modeled and analyzed by these approaches.

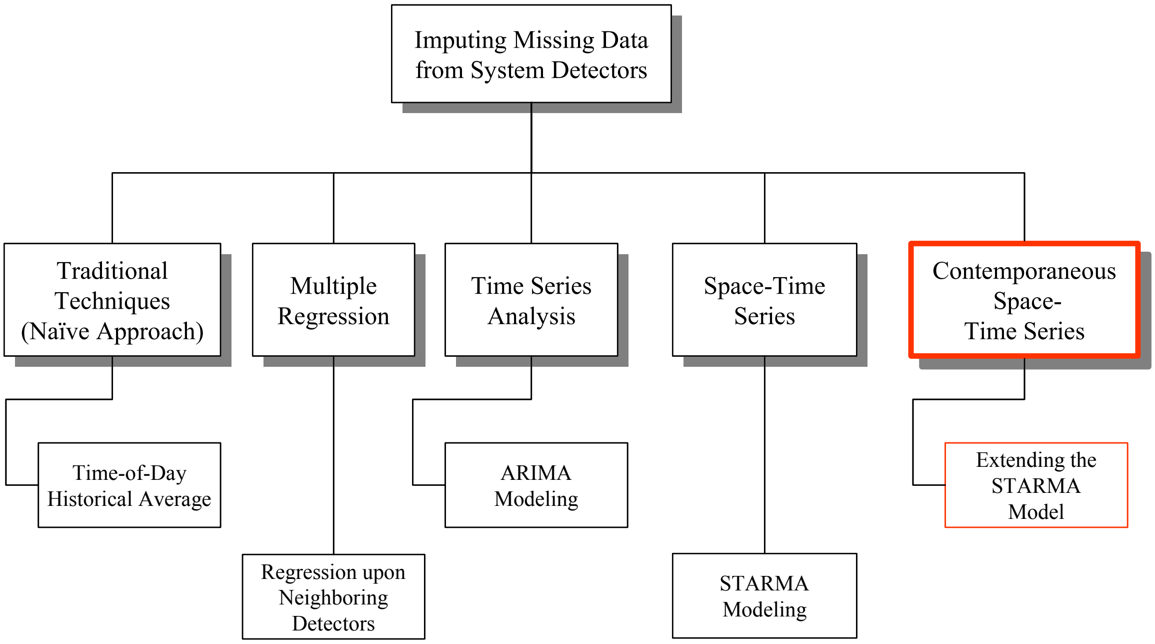

Imputing Missing Data from System Detectors
Traditional Techniques (Naïve Approach)
Multiple Regression
Time Series Analysis
Space-Time Series
Contemporaneous Space-Time Series
Time-of-Day Historical Average
Regression upon Neighboring Detectors
ARIMA Modeling
STARMA Modeling
Extending the STARMA Model

Figure 3-1: Techniques to Impute Missing Data

3.2.1 Time-of-Day (TOD) Historical Average

Mean imputation substitutes the mean value of the available data for the missing data values. This is a simplistic method that replaces the missing value(s) using the statistical mean of the available data. The general model to derive the sample mean, or historical average, is as follows:

The sample mean x(bar) of a set of numbers $x_1, x_2, \ldots, x_n$ *is given by:*

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

In the domain of traffic data, our hypothesis is to substitute the mean, or historical average, of the time-of-day interval for which missing data is observed. For example, if a detector fails to supply volume data for the 10:00-10:15 AM interval, the historical average of vehicular volume for that particular interval could be used to feed the signal control system. The following example illustrates how the historical average volume would be substituted for the time-of-day interval.

Table 3-3: Sample calculation of historical average model

Sample calculation of Historical Average

	VOLUME					
TIME OF DAY	**Day 1**	**Day 2**	**Day 3**	**Day 4**	**Day 5**	**AVERAGE VOLUME**
10:00	312	364	360	400	416	370.4
10:15	356	352	340	312	440	360.0
10:30	296	272	352	364	284	313.6
10:45	224	344	292	228	360	289.6
11:00	248	304	276	276	280	276.8
11:15	336	236	448	288	312	324.0
11:30	264	396	364	388	316	345.6
11:45	364	356	380	356	376	366.4
12:00	240	336	312	388	324	320.0

3.2.2 Multiple Regression on Neighboring Detectors

Regression imputation estimates the missing values by regression of the variable of interest on the other variables. The general linear model is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_k x_k + \varepsilon$$

where,

y	is the dependent variable
$x_1, x_2, \ldots, x_k$	are the independent variables
$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_k x_k$	is the deterministic portion of the model
ε	is the probabilistic term

Regression enables us to discover any underlying relationships between the detector of interest and its neighboring detectors. The hypothesis is that we can impute the missing traffic data at a particular detector by regressing upon nearby or spatially correlated detectors within the same arterial network. Since we are working with fifteen-minute traffic data, we would be able to explore the effects of upstream and downstream traffic flow on imputing missing values at a specific intersection.

The following example illustrates the implementation of the general linear regression model. Referring to the Figure 3-2 below, suppose we wish to regress the 15-minute volume for detector 2027 upon other detectors within the arterial network such as detectors 2008, 2021, 2022, and 2031. (A feature selection procedure should be performed to determine which detectors are suitable independent variables in the regression model.) The resulting regression model to impute the volume at any time interval *t* for detector 2027 would be:

$$\hat{V}(t)_{Det.2027} = \beta_0 + \beta_1 V(t)_{Det.2008} + \beta_2 V(t)_{Det.2021} + \beta_3 V(t)_{Det.2022} + \beta_4 V(t)_{Det.2031}$$

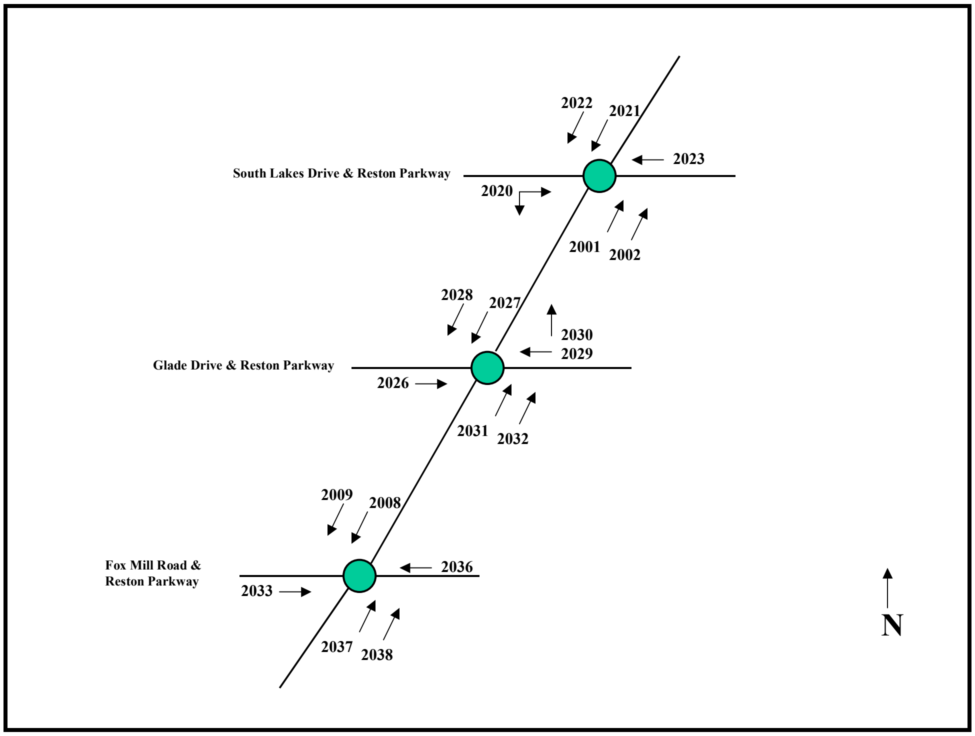


Figure 3-2: System detectors along the Reston Parkway (Reston, VA)

3.2.3 Time Series Analysis

Time Series analysis (ARIMA) estimates missing data using data that preceded the missing values in time order. Traffic data is collected at regular intervals throughout the day and, therefore, can be readily modeled using time series analysis techniques. Williams et al. (1999) have shown that seasonal autoregressive moving average (Seasonal ARIMA) models perform well at forecasting 15-minute traffic data (volume and occupancy). We wish to apply a suitable seasonal ARIMA model as supported in their research to estimate traffic data at any 15-minute time interval for which missing data occurs. Thus, the data leading up to the interval of missing data would be used to estimate, or forecast, the value for that interval.

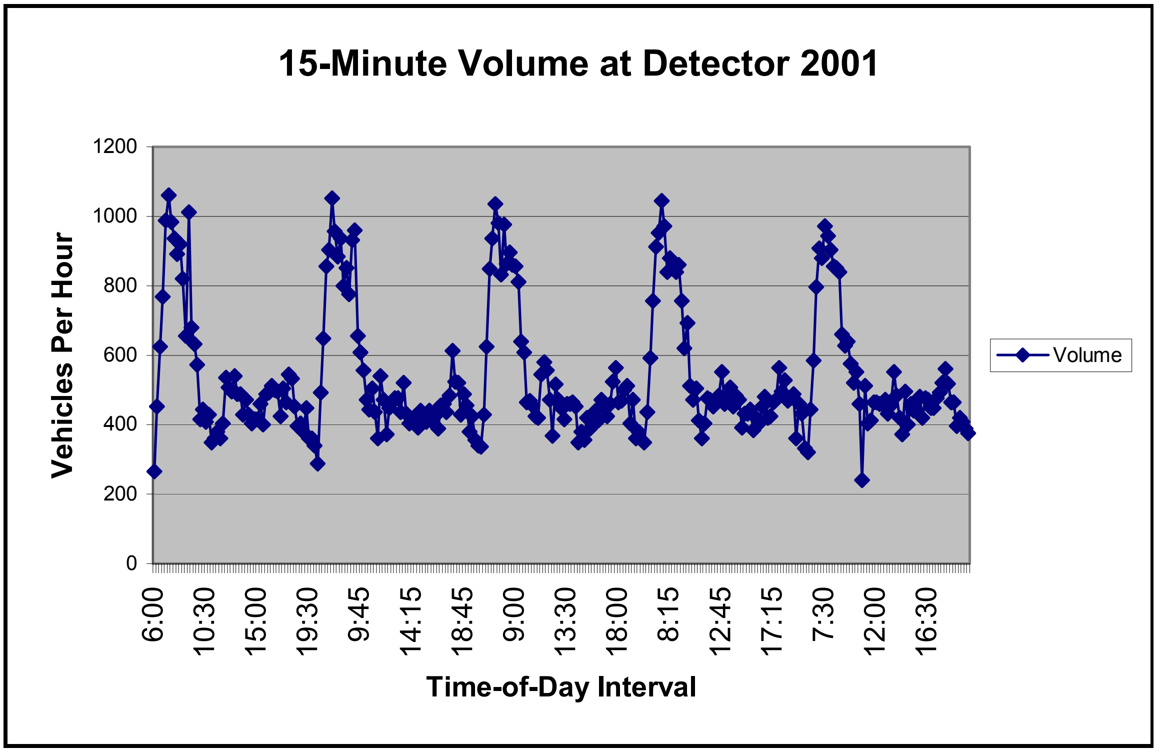


Figure 3-3: Time series plot of volume data

3.2.4 Space-Time Autoregressive Moving Average (STARMA)

Analogous to univariate time series, Space-Time Autoregressive Moving Average (STARMA) models can be expressed as a linear combination of past observations and errors. However, instead of allowing dependence of forecasted values on only with past observations and errors at one location, dependence is allowed with neighboring locations of various spatial orders (Pfeifer, Deutsch 1980). The STARMA is an extension of the univariate time-series model that takes into account time-series data from neighboring locations, thus potentially improving the forecasts of future values at a particular location. In our efforts to estimate missing data at a specified system detector, the STARMA model takes into account the time series data at this system detector, as well as time series data from neighboring detectors.

The STARMA model is presented in a series of papers by Pfeifer and Deutsch (Pfeifer Deutsch 1980). These papers discuss the theoretical foundation of the STARMA model, as well as model implementation procedures, and are as follows: the STARMA($p_{\lambda 1,\ \lambda 2,\ ..,\ \lambda p}$, $q_{m1,\ m2,\ ..,\ mq}$) model class is characterized by linear dependence lagged in both space and time. The autoregressive form of the space-time model would express the observation at time t and site i, $z_i(t)$ as a linear combination of past observations at site i and neighboring sites.

Notation:

$z_i(t)$	Observation of the random variable $Z_i(t)$
t	Time index ($t = 1, .., T$)
N	Number of fixed sites in space ($i = 1, 2, .., N$)
$L^{(l)}$	Spatial lag operator of spatial order l

The STARMA model class is expressed as follows:

$$z_i(t)=\sum_{k=1}^{p}\sum_{l=0}^{\lambda_k}\phi_{kl}W^{(l)}z_i(t-k)+\varepsilon_i(t)-\sum_{k=1}^{q}\sum_{l=0}^{m_k}\theta_{kl}W^{(l)}\varepsilon_i(t-k)$$

where,

p	is the autoregressive order,
q	is the moving average order,
λ_k	is the spatial order of the k^{th} autoregressive term,
m_k	is the spatial order of the k^{th} moving-average term
ϕ_{kl}	is the autoregressive parameter at temporal lag k and spatial lag l,
θ_{kl}	is the moving-average parameter at temporal lag k and spatial lag l,
$W^{(l)}$	is the N x N matrix of weights for spatial order l, and

$\varepsilon(t)$ is the random normally distributed error vector at time t

3.2.4.1 Procedure for Space-Time Modeling:

1.0 Identification of STARMA models:

The first step is to determine which of the model forms (STAR, STMA, STARMA) is the most appropriate for the data at hand, and its associated temporal and spatial orders (p, q, λ, m). The purpose of the identification process is to choose the model class that exhibits theoretical properties that most closely matches those estimated from the data. In univariate time series analysis, the primary tools in identification are the autocorrelation and partial autocorrelation functions. Choosing between the three general subclasses of models (AR, MA, ARMA) is a matter of determining whether the partial autocorrelation functions cuts off, the autocorrelation function cuts off, or they both tail dissipate.

In space-time models, Pfeifer proposes to combine the N^2 possible cross-covariances between all possible pairs of sites in a logical manner consistent with the forms associated with the selected model class. This is referred to as the space-time autocovariance function, which expresses the covariance between points lagged in both space and time. An average covariance between the weighted l^{th} order neighbors of any site and the k^{th} order neighbors of the same site at s time lags in the future would be:

$$\gamma_{lk}(s) = E\left[\sum_{i=1}^{N} \frac{L^{(l)} z_i(t)\, L^{(k)} z_i(t+s)}{N}\right]$$

where $\gamma_{lk}(s)$ is the space-time autocovariance between l^{th} and k^{th} order neighbors at time lag s. The sample estimate of the space-time autocovariance function is:

$$\hat{\gamma}_{lk}(s) = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T-s} L^{(l)} z_i(t)\, L^{(k)} z_i(t+s)}{N(T-s)}$$

This leads to the sample estimate of the **space-time autocorrelation function** between l^{th} and k^{th} order neighbors at s time lags apart. This function approximates constant variance at all spatial lags.

$$\hat{\rho}_{lk}(s) = \frac{\hat{\gamma}_{lk}(s)}{\left[\hat{\gamma}_{ll}(0)\,\hat{\gamma}_{kk}(0)\right]^{1/2}}$$

$$= \frac{\sum_{i=1}^{N}\sum_{t=1}^{T-s} L^{(l)} z_i(t)\, L^{(k)} z_i(t+s)}{\left[\sum_{i=1}^{N}\sum_{t=1}^{T} (L^{(l)} z_i(t))^2 \cdot \sum_{i=1}^{N}\sum_{t=1}^{T} (L^{(k)} z_i(t))^2\right]}$$

The **space-time partial autocorrelation function** is defined as:

$$\gamma_{h0}(s) = \sum_{j=1}^{k}\sum_{l=0}^{\lambda} \phi_{jl}\gamma_{hl}(s-j)$$

Analogous to univariate time series, STARMA processes are characterized by a distinct space-time partial and autocorrelation function. The characteristics of the theoretical space-time autocorrelation functions are as follows:

Table 3-4: Identifying the STARMA Model Type

Model Form	**Space-Time Autocorrelation Function**	**Space-Time Partial Autocorrelation Function**
STAR($p_{\lambda 1..\ \lambda p}$)	Tails off	Cuts off after p time lags, λ_p spatial lags
STMA($q_{m1..\ mp}$)	Cuts of after q time lags, m_q spatial lags	Tails off
STARMA(p, q)	Tails off	Tails off

2. Estimation of the STARMA model:

After a candidate model form has been selected, the next phase of the modeling procedure is to estimate the ϕ and θ parameters. Pfeifer argues that techniques based on standard linear regression theory are suitable for estimation of STAR model parameters. As an example, consider the STAR(2_{10}) model:

$$z(t) = \phi_{10} z(t-1) + \phi_{11} W^{(1)} z(t-1) + \phi_{20} z(t-2) + \varepsilon(t)$$

In general linear model form **Y = XB +** ε, this model for *t = 1, 2, .., T* can be written as follows:

$$\begin{bmatrix} z(1) \\ z(2) \\ z(3) \\ \vdots \\ z(T) \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ z(1) & W^{(1)} z(1) & 0 \\ z(2) & W^{(1)} z(2) & z(1) \\ \vdots & \vdots & \vdots \\ z(T-1) & W^{(1)} z(T-1) & z(T-2) \end{bmatrix} \begin{bmatrix} \phi_{10} \\ \phi_{11} \\ \phi_{20} \end{bmatrix} + \begin{bmatrix} \varepsilon(1) \\ \varepsilon(2) \\ \varepsilon(3) \\ \vdots \\ \varepsilon(T) \end{bmatrix}$$

The zero vectors were substituted for the unobserved z vectors, for those times before the system was under observation. It should be noted that due to the time series nature of the STAR model, the linear regression assumptions about the independent or regressor variable do not hold. Specifically, the X matrix is stochastic rather than deterministic in repeated samples. Also, the residuals are not normally independently distributed with mean zero and constant variance. The linear regression results will be used here with apriori knowledge that they are only approximate. Due to the non-linear form of the STMA and STARMA models, non-linear optimization techniques must be employed to estimate the associated parameters.

3. Diagnostic Checking of the STARMA model:

After a candidate model has been selected and its parameters estimated, the model must undergo evaluation to determine whether the model adequately represents the data. The first phase is to perform an analysis of the residuals: if the fitted model is adequate, the residuals should be white noise, i.e. variance-covariance matrix equal to $\sigma^2 I_N$ and all autocovariances at non-zero lags equal to 0. The second phase is to verify the statistical significance of the estimated parameters.

3.2.4.2 Example STARMA Model

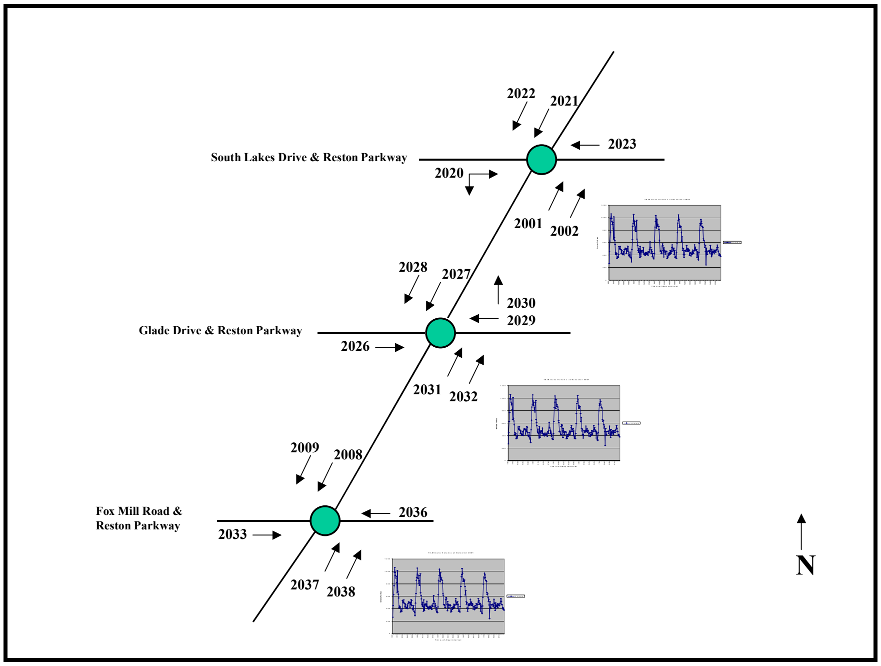


Figure 3-4: Example layout of STARMA model

As example of a STARMA model of spatial order 1, suppose we wish to estimate volume for detector 2032 in Figure 3-4, using detectors 2002 and 2038 as input data sources. A STARMA (1, 1) model indicates an order of one for both autoregressive and moving average parameters, as well as a spatial order of one.

$$\begin{aligned}\hat{V}(t)_{Det.2032} = {} & \phi_{10}V(t-1)_{Det.2032} + \phi_{11}V(t-1)_{Det.2002} + \phi_{11}V(t-1)_{Det.2038} \\ & + \theta_{10}V(t-1)_{Det.2032} + \theta_{11}V(t-1)_{Det.2002} + \theta_{11}V(t-1)_{Det.2038} \\ & + \varepsilon(t)\end{aligned}$$

where, ϕ_{kl} = autoregressive parameter at temporal lag *k* and spatial lag *l*

θ_{kl} = moving average parameter at temporal lag *k* and spatial lag *l*

3.2.5 Group Model

The techniques that we apply to estimate missing data will derive different results primarily due to how they apply available system detector data. Finally, we consider a model that takes each of these individual estimations into account similar to that of a consensus model. This group model comprises of a weighted combination of the previously explored models' estimations of missing system detector data. A simple approach to constructing this model is a simple linear (multiple) regression upon the models' estimates of traffic data.

$$\begin{aligned}\hat{V}(t) = {} & \beta_0 + \beta_1 V(t)_{Historical\ Average} + \beta_2 V(t)_{\mathrm{Re}gression} + \beta_3 V(t)_{Time\ Series} \\ & + \beta_4 V(t)_{STARMA} + \beta_5 V(t)_{UniqueSTARMA}\end{aligned}$$

4.0 PROBLEM FORMULATION & IMPLEMENTATION

4.1 System Detector Data

Traffic data are collected by system detectors which are embedded on each approach to the intersections. Detector data is then archived at the NOVA Operations Center and also fed to the systems at the STL. This research was performed using system detector from arterial networks at the individual detector level (vice station level). System detector data at the station level is very effective at describing the current traffic conditions, however, this research addresses the worst-case scenario where individual detectors are off-line.

4.1.1 Data Selection: Three-Intersection Case

The scope of our research in estimating missing traffic data is to derive estimations of vehicular volume (vehicles per hour) and occupancy (percent of hour). The Virginia Department of Transportation's (VDOT) Smart Traffic Signal Systems (STSS) division provided the necessary data to conduct our research. We shall focus upon the Reston Area Network (RAN) of coordinated intersections that is situated in suburbs of northern Virginia. The main throughway within the RAN is the Reston Parkway.

Traffic data has been collected for this arterial network via single-loop inductance wires and archived at the University of Virginia's Smart Travel Lab databases. Historical data spans from January 2000 to the present. We selected a series of intersections along the Reston Parkway for which there was sufficient data to perform our model development and analysis.

We focus our model development efforts on three consecutive intersections on the Reston Parkway. From North to South, this sub-network comprises the intersections of Reston Parkway and South Lakes Drive (Figure 4-1), Glade Drive (Figure 4-2), and Fox Mill Road (Figure 4-3). System detectors are situated along each of the four approaches to each intersection. This sub-network is graphically represented by Figure 4-4 and each individual system detector is identified with a unique identifier number. The arrows in Figure 4.4 indicate the traffic movement for which that particular system detector collects data. For instance, detector 2001 collects volume and occupancy data for a single-lane, northbound movement at the intersection of Reston Parkway & South Lakes Drive.

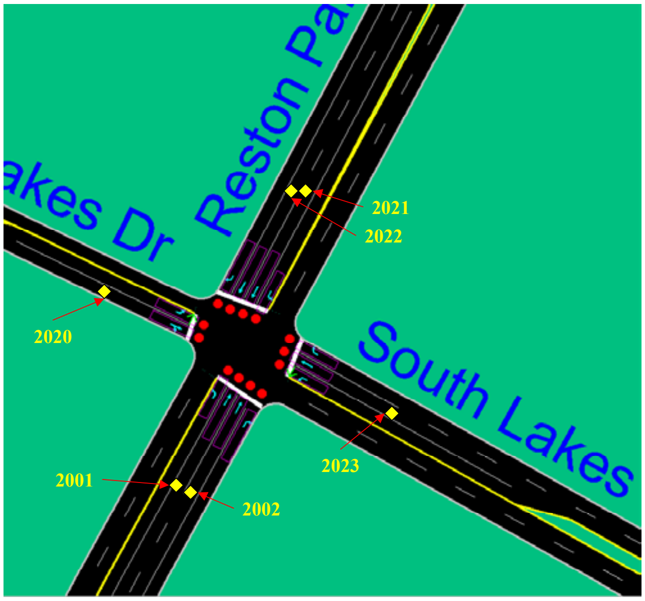


Figure 4-1: Intersection of Reston Parkway & South Lakes Drive

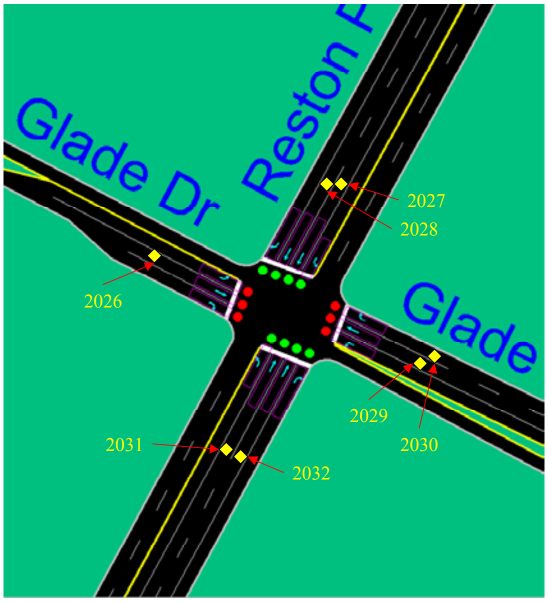


Figure 4-2: Intersection of Reston Parkway & Glade Drive

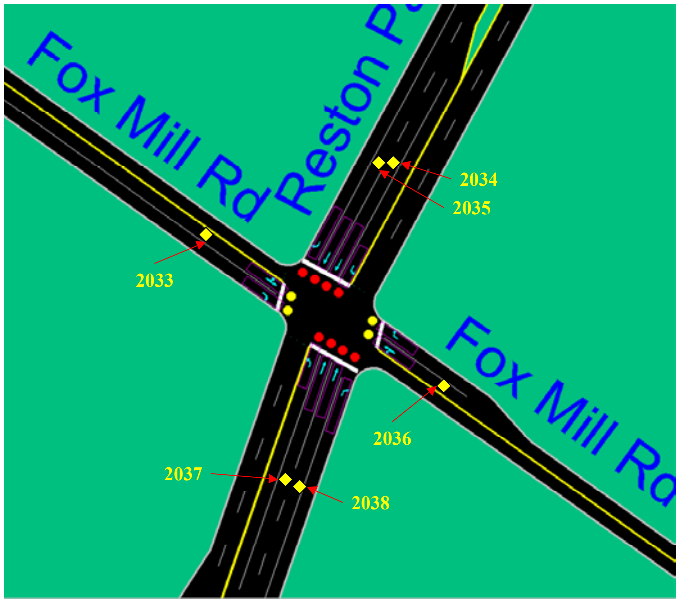


Figure 4-3: Intersection of Reston Parkway & Fox Mill Road

Figure 4-4 represents the three-intersection sub-network of the Reston corridor. Each of the available system detectors is labeled and indicates the approach path for which each detector collects traffic volume and occupancy data. (Note: diagram is not drawn to scale and does not convey exact locations of system detectors.)

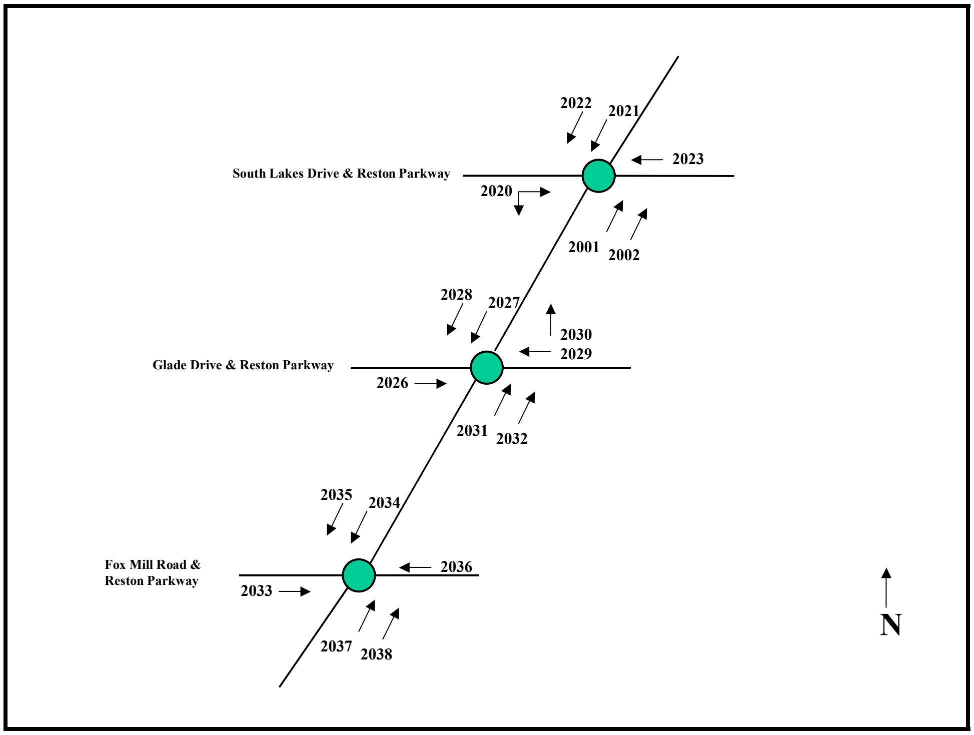


Figure 4-4: Three-Intersection Sub-Network of the Reston Parkway

4.1.2 Data Selection: Corridor Network

In addition to the investigating model performance on a small sub-network scale, the models were also implemented against a larger data set. We subjected our model building procedures to include more than 40 detectors from the Reston Area Network, instead of using only the eighteen surrounding detectors as input sources (as in the previous scenarios). The availability of a larger set of data sources better simulates the conditions encountered in real-world traffic operations.

4.1.3 Descriptive Statistics on System Detector Data

Although system detector data is available for 24-hours per day, we specified the input data for our model development efforts to span from 6:00 AM – 8:00 PM. This 14-hour period exhibits the typical business day traffic demand and provides sufficient data for our analysis. The training set used to develop the various models consisted of system detector data spanning March-August 2000 from the Reston Area Network. The test or evaluation set consisted of ten day's worth of system detector data.

System detector data is continuously archived by the University of Virginia's Smart Travel Lab in fifteen-minute intervals. For this 14-hour period, fifty-six time intervals, or observations, per day. The following plots (Figure 4-5) illustrate the average daily volume and average daily occupancy for each system detector of the three intersections for northbound traffic (single lane – detector 2001) at the intersection of Reston Parkway & South Lakes Drive.

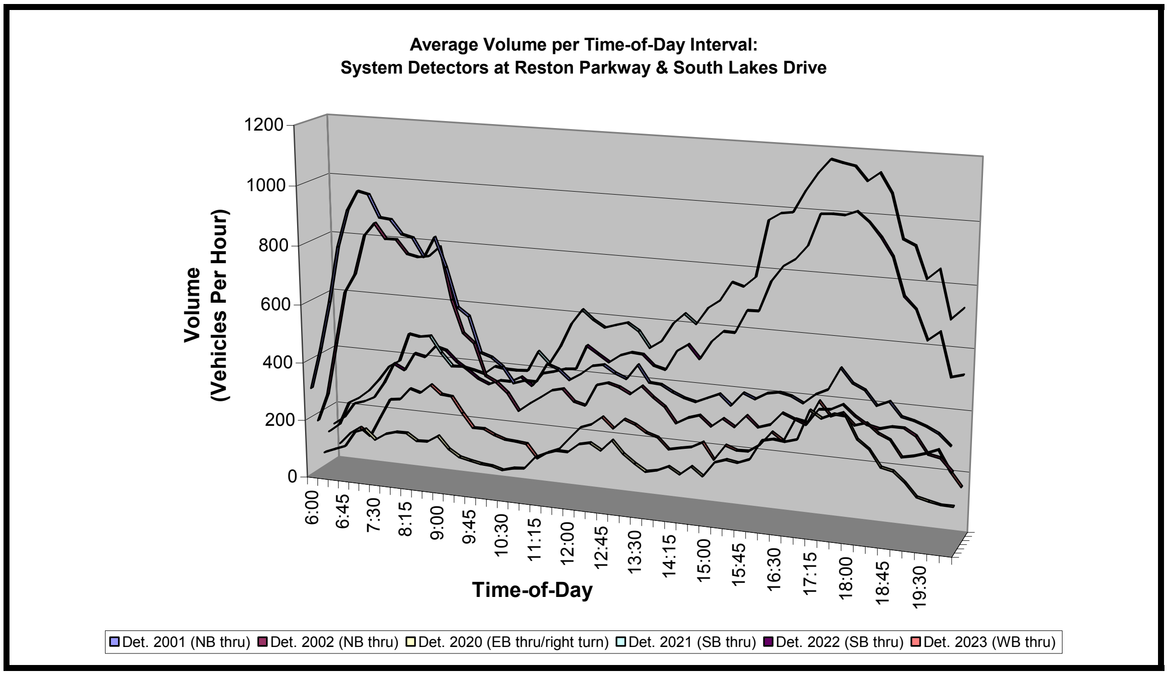


Figure 4-5: Average Volume vs. TOD at Reston Parkway & South Lakes Drive

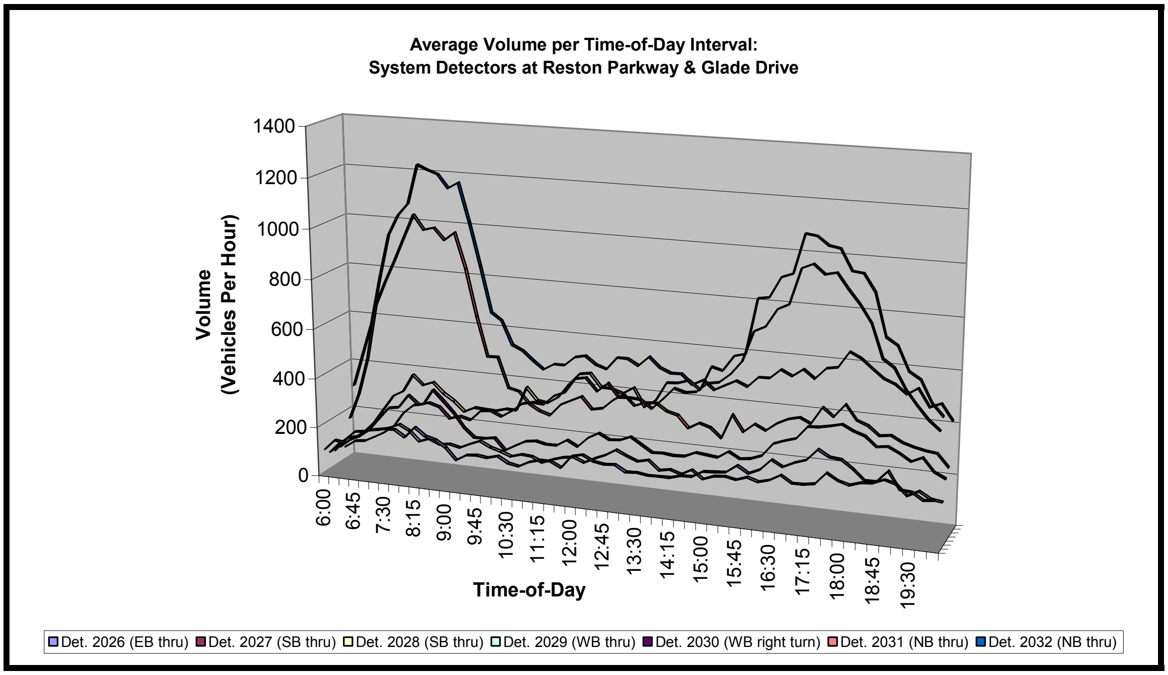


Figure 4-6: Average Volume vs. TOD at Reston Parkway & Glade Drive

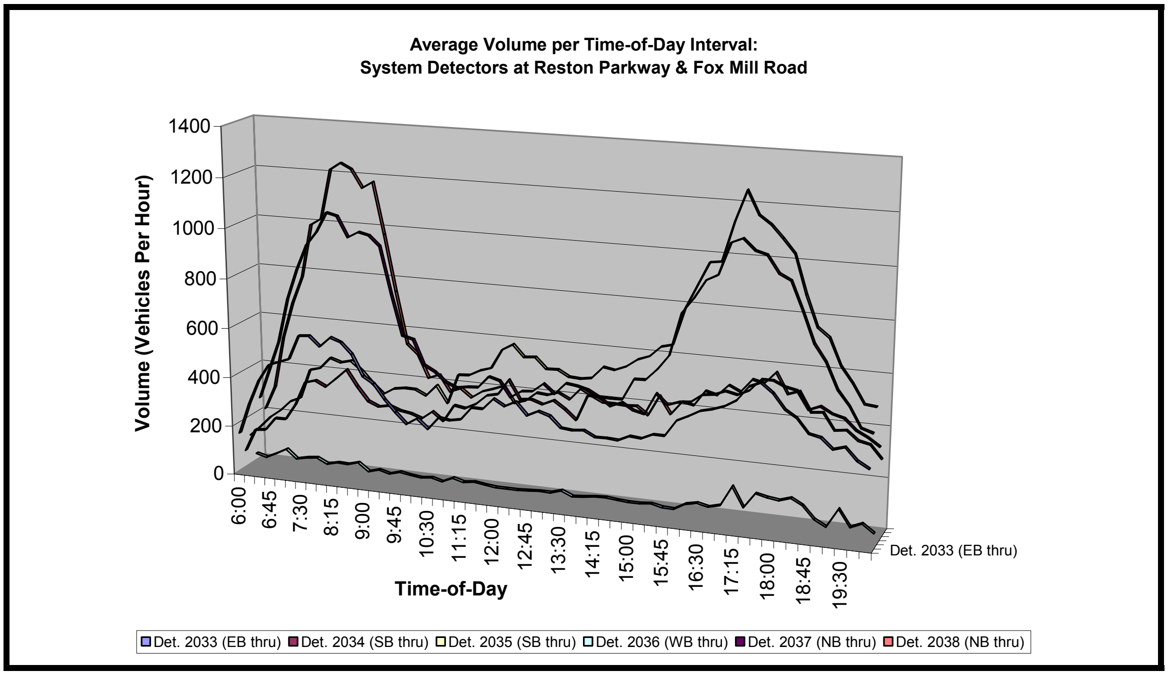


Figure 4-7: Average Volume vs. TOD at Reston Parkway & Fox Mill Road

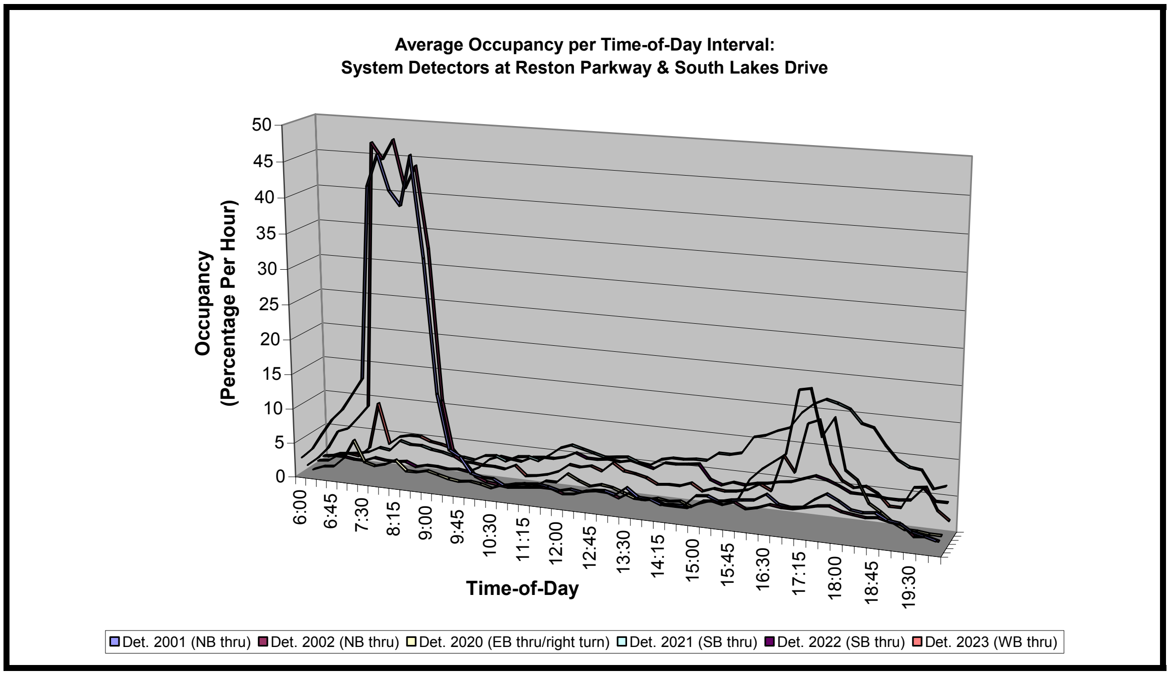


Figure 4-8: Average Occupancy vs. TOD at Reston Parkway & South Lakes Drive

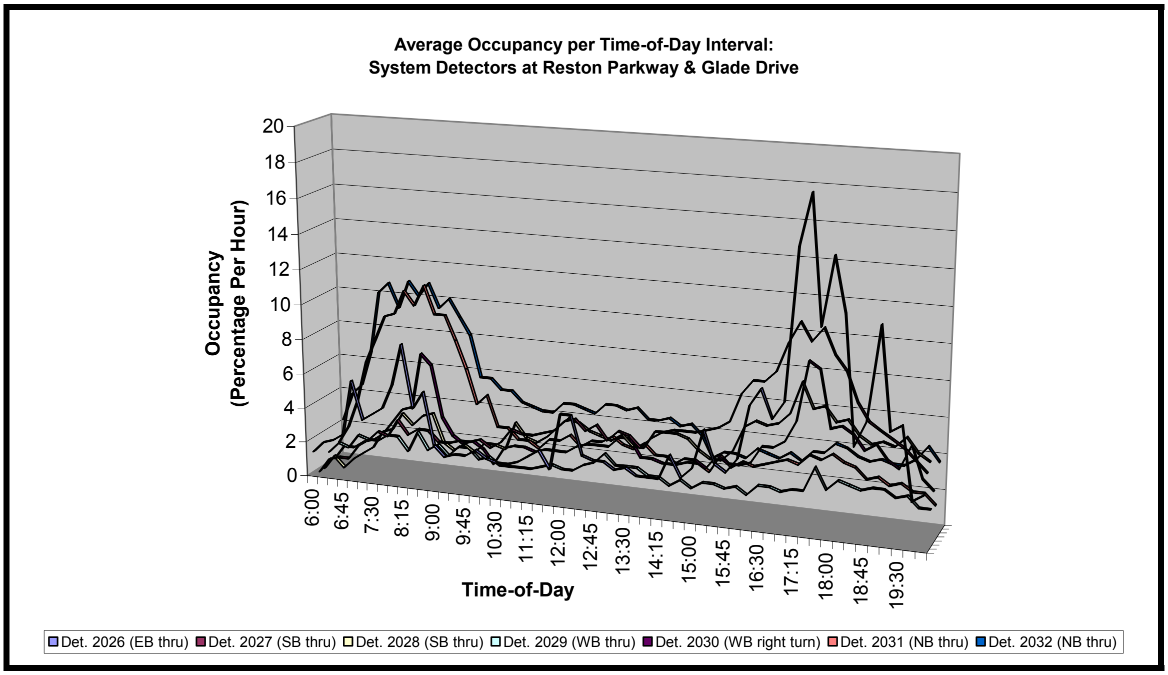


Figure 4-9: Average Occupancy vs. TOD at Reston Parkway & Glade Drive

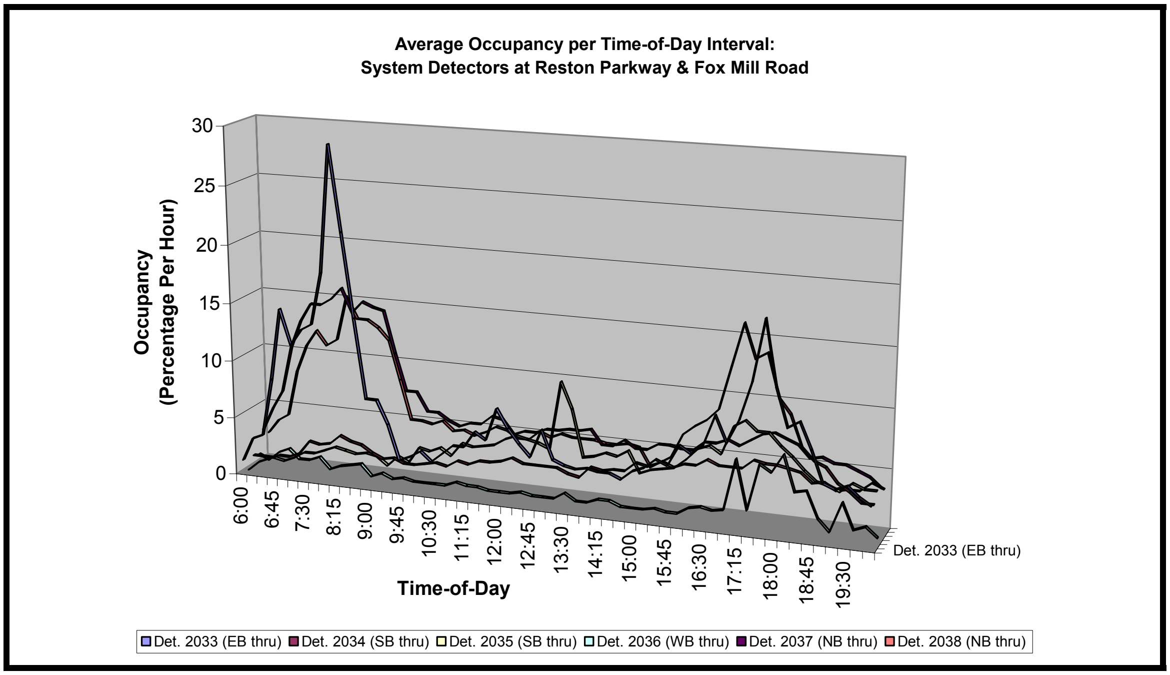


Figure 4-10: Average Occupancy vs. TOD at Reston Parkway & Fox Mill Road

4.2 Model Scenarios

We shall examine three distinct network geometries for each of the modeling approaches. The network geometries are representative scenarios of the availability of data sources to our estimation models. We shall examine the performance of the relevant models to reflect these scenarios. The testing procedure called for imputing traffic data (volume and occupancy) for the selected detector at each time-of-day interval, and then comparing with the actual values.

4.2.1 Models Using Upstream Detectors Only: Detector 2001

This geometry depicts the scenario when only upstream detectors, in addition to those from the location-of-interest, are available as data sources to estimate traffic parameters at a desired point. Furthermore, assume that the only additional data sources are detectors from upstream intersections, namely the South Lakes Drive and Glade Drive intersections. Therefore, this model would use upstream data sources to estimate traffic parameters at a downstream location.

1. Detector 2001: Northbound detector at South Lakes Drive & Reston Parkway

Detector 2001 is the system detector located in the inside-most lane at this intersection. It captures volume and occupancy data for the single-northbound lane on the mainline. This detector shall be the focus for the model using only upstream detectors as input data.

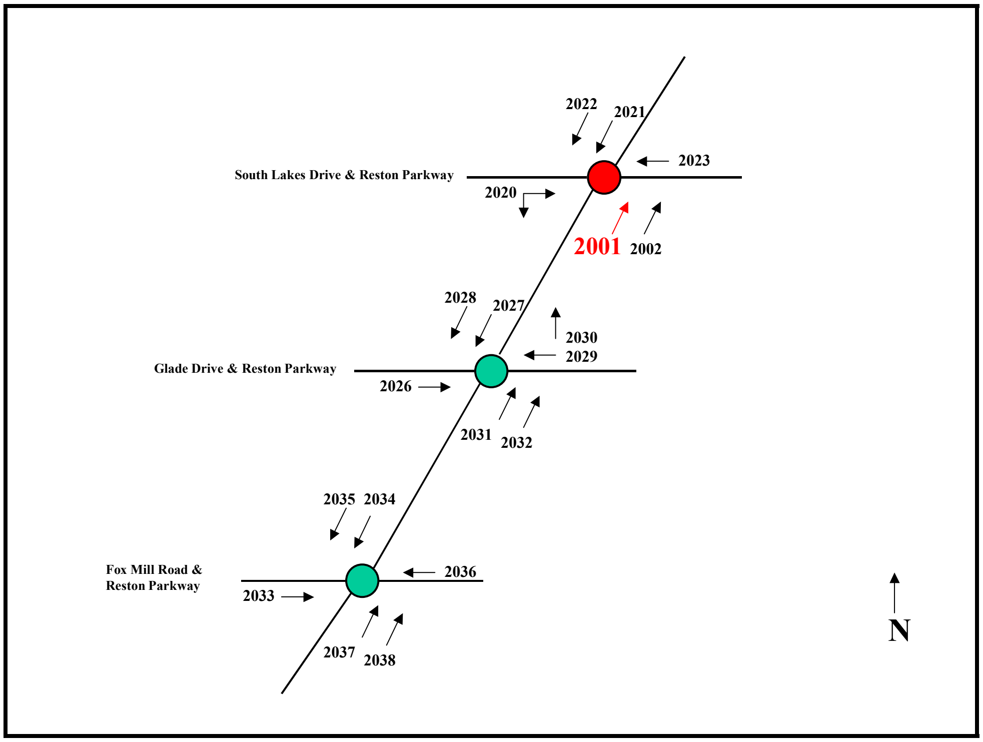


Figure 4-11: Model using data from upstream detectors only

4.2.2 Models Using Both Upstream and Downstream Detectors: Detector 2027

This geometry depicts the scenario when both upstream and downstream detectors, in addition to those from the location-of-interest, are available as data sources to estimate traffic parameters at a desired intersection. The detectors from the upstream intersection, South Lakes Drive, and the downstream intersection, Fox Mill Road, provide the input data for the various models. Therefore, this model uses data from both upstream and downstream sources to estimate traffic parameters at the particular location.

2. Detector 2027: Southbound detector at Glade Drive & Reston Parkway

Detector 2027 is the system detector located in the inside-most lane at this intersection. It captures volume and occupancy data for the single-southbound lane on the mainline. This detector shall be the focus for the model using both upstream and downstream detectors as input data.

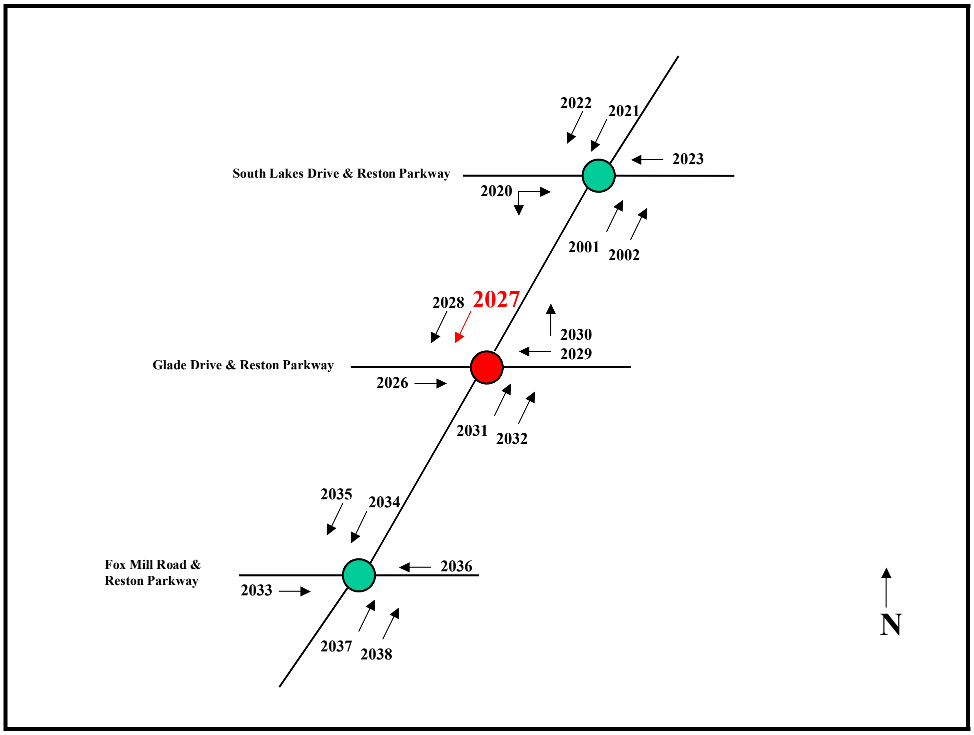

2022
2021
2023
South Lakes Drive & Reston Parkway
2020
2001
2002
2028
2027
2030
2029
Glade Drive & Reston Parkway
2026
2031
2032
2035
2034
Fox Mill Road & Reston Parkway
2036
2033
2037
2038
N

Figure 4-12: Model using data from both upstream and downstream detectors

4.2.3 Models Using Downstream Detectors Only: Detector 2037

This geometry depicts the scenario when only downstream detectors, in addition to those from the location-of-interest, are available as data sources to estimate traffic parameters at a desired intersection. Furthermore, assume that the only additional data sources are detectors from downstream intersections, namely the Glade Drive and Fox

Mill Road intersections. Therefore, this model would use only downstream data sources to estimate traffic parameters at a downstream location.

3. Detector 2037: Northbound detector at Fox Mill Road & Reston Parkway

Detector 2037 is the system detector located in the inside-most lane at this intersection. It captures volume and occupancy data for the single-northbound lane on the parkway. This detector shall be the focus for the model using only downstream detectors as input data.

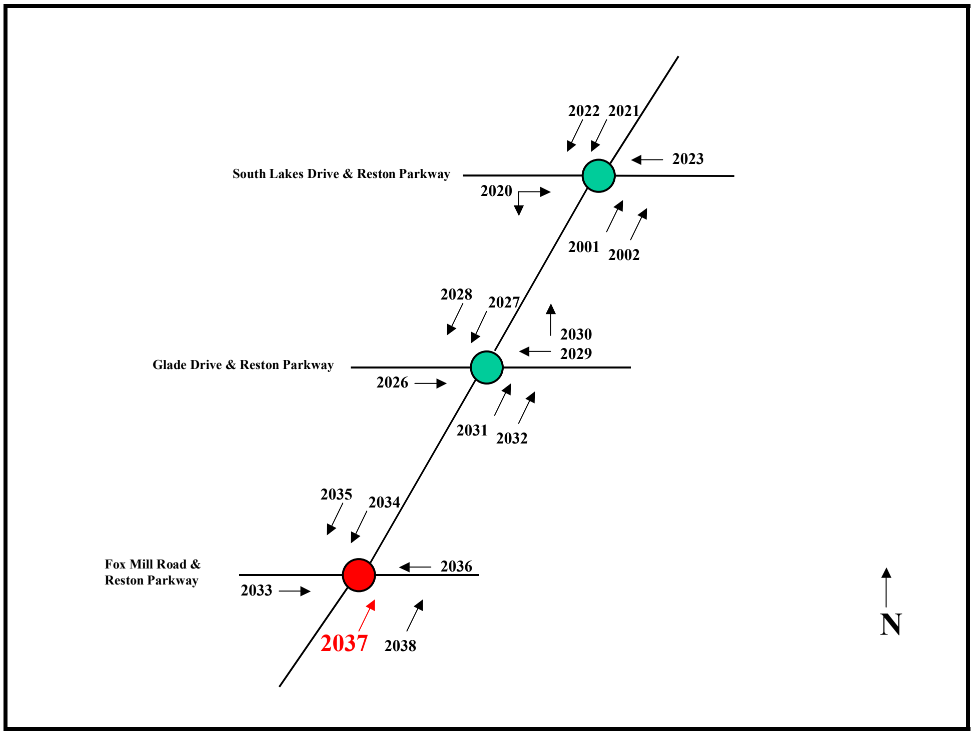


Figure 4-13: Model using data from downstream detectors only

4.2.4 Models Using Extended Network Data

These models were developed using a significantly larger data than the previous scenarios. We again revisited imputing data for detector 2001; however, the models were exposed to more input data sources than the previous models. Instead of using only the eighteen surrounding detectors as input sources to detector 2001 models (as in the previous case), we subject our model building procedures to include more than 40 additional detectors from the Reston Area Network. Figure 4-14 illustrates the extended network.

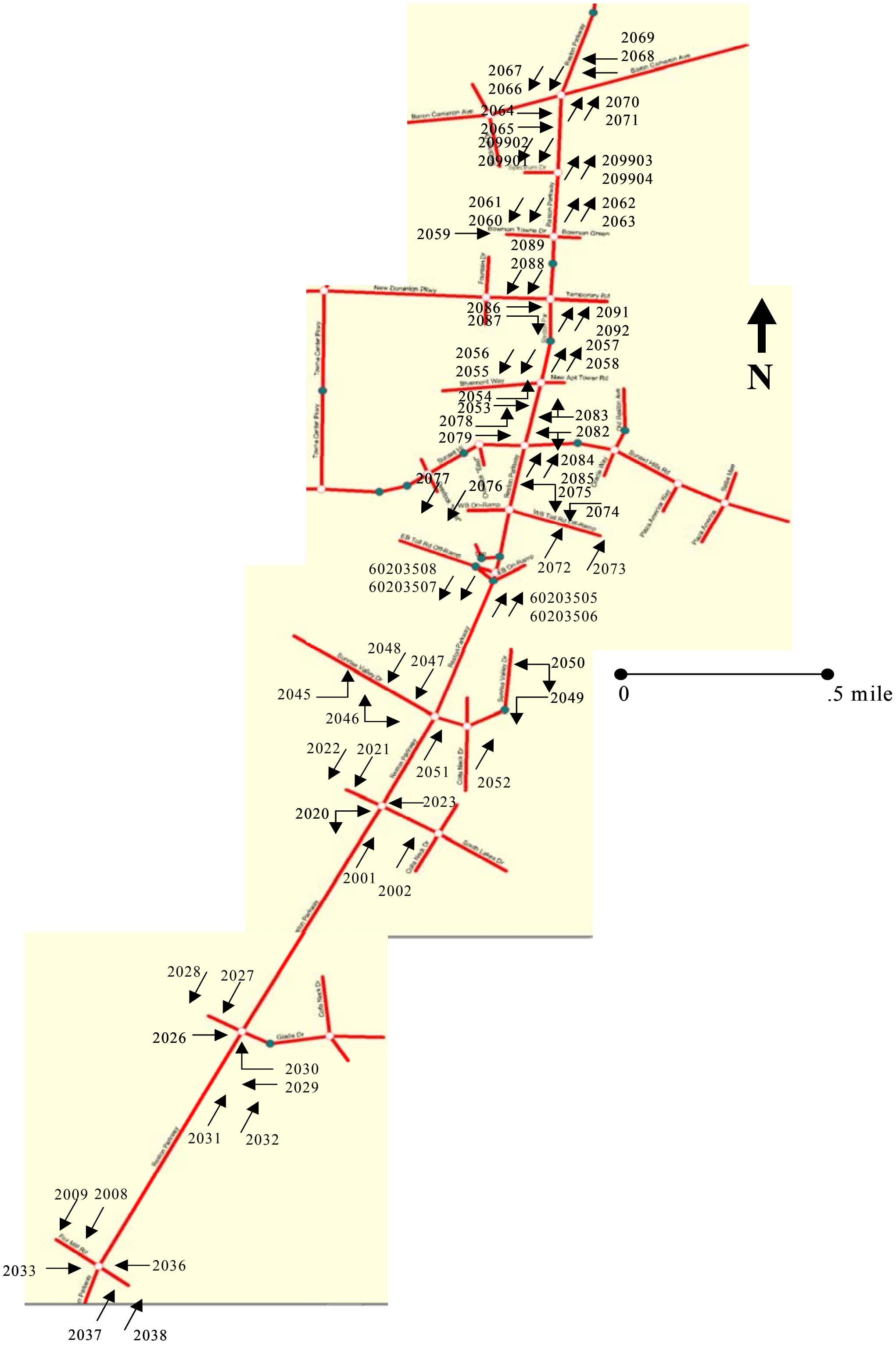


Figure 4-14: Extended Network Scenario

4.3 Model Development

4.3.1 Historical Average Model

This naïve approach is the baseline from which to compare the performance of the remaining models. Most first-generation traffic signal systems, such as PB Faradyne's MIST, calculate the historical averages of volume and occupancy for each 15-minute interval at each system detector throughout the day. Therefore, this approach is readily applicable in practice. We wish to validate the effectiveness of this simple approach to estimating traffic data. For each of the detectors we specified to model, this model calculates the historical average volume and occupancy at each time-of-day interval within the 14-hour period.

The model development procedure is simply to calculate the historical average volume and occupancy for each time-of-day interval within period of interest. This estimation model applies to all three network geometries but does not require development of distinct models.

4.3.2 Regression Model

These models estimate a detector's volume or occupancy for a particular time interval by regressing upon other detectors' volume or occupancy values at the same time interval. We developed distinct models for each of the three network geometries. Depending upon the network geometry, different detectors may be selected as predictor variables during the feature selection step. In addition, separate models are developed to estimate data for volume and occupancy.

The model estimation procedure is as follows:

Step 1: Perform feature selection to determine the critical predictor variables within the sub-network to the dependent variable, i.e. the system detector of interest. We choose to use the *Mallow's* C_p criterion to select the input variables.

Step 2: Fit a multiple linear regression model using the selected detectors' data

Step 3: Perform tests to verify that the model's residuals are approximately NID $(0, \sigma^2)$

Step 4: Check the utility of the model according to *Adjusted-*R^2 criterion

Step 5: If the model is adequate, then estimate traffic data (volume or occupancy) for each time-of-day interval within period of interest

4.3.3 Time Series (ARIMA) Model

These models estimate a detector's volume or occupancy for a particular time interval by treating the traffic data as a time series. Separate time series models are developed for both volume and occupancy. Williams et al (1999) investigated the applicability of seasonal ARIMA models to traffic data; we shall apply their findings in our model development.

The time series model estimation procedure is as follows:

Step 1: Plot traffic data (volume or occupancy) against time-of-day interval

Step 2: Develop sample autocorrelation (ACF) and partial-autocorrelation (PACF) correlograms. Determine seasonal and non-seasonal components and periodicity.

Step 3: Fit appropriate Seasonal ARIMA model

Step 4: Perform tests for randomness of model's residuals

Step 5: If the model is adequate, then estimate traffic data (volume or occupancy) for each time-of-day interval within period of interest

4.3.4 Space-Time Series Model (STARMA)

As described previously, these models extend the ARMA univariate time series models into the spatial domain by taking into consideration time series data from neighboring locations to the point-of-interest. In our study, time series data from neighboring detectors can be used in conjunction with time series data at the detector-of-interest to estimate traffic data for that location. Again, separate models are constructed to estimate volume and occupancy values.

To simplify the construction of space-time series models, we make the following assumptions: Neighboring detectors shall be equally weighted; this is analogous to the STARMA spatial order of one designation. We shall assume that model parameter estimation, such as linear regression techniques, can substitute for the weighting matrix.

We shall also substitute examining cross-correlation functions for STARMA autocorrelation and partial autocorrelation functions. This is a reasonable approach for our study since cross-correlation functions exhibit correlation between multiple time series. The observations of one series are correlated with the observations of another series at various lags and leads. In addition, commercial statistical applications, such as SPSS, include functions to develop these models.

Details of the STARMA model building procedure were presented in Section 3.2.4. Generally, the spatial-time series model estimation procedure is as follows:

Step 1: Plot traffic data (volume or occupancy) against time-of-day interval

Step 2: Develop sample autocorrelation (ACF) and partial-autocorrelation (PACF) correlograms; alternatively, develop cross-correlation functions to determine correlations between time series data at the detector-of-interest and neighboring detectors.

Step 3: Fit appropriate space-time series model

Step 4: Perform tests for randomness of model's residuals

Step 5: If the model is adequate, then estimate traffic data (volume or occupancy) for each time-of-day interval within period of interest

4.3.5 Group Model

This model fits a multiple regression model to the estimates derived by the previous models. This weighted model simulates a consensus model by determining if multiple estimations serve well to determine traffic data at time interval *t*. The combined regression & spatial-time series model estimation procedure is as follows:

Step 1: Fit a multiple linear regression model to the previous models' estimations of volume and occupancy

Step 2: Perform tests to verify that the model's residuals are approximately NID $(0, \sigma^2)$

Step 3: Check the utility of the model according to *Adjusted-*R^2 criteria

Step 4: If the model is adequate, then estimate traffic data (volume or occupancy) for each time-of-day interval within period of interest

4.4 Model Execution

Historical averages for each time-of-day interval were derived from ITS databases. Regression and weighted models were executed using Minitab v13.0. Time Series and STARMA models were executed using SPSS v10.3, which include functionality to approximate STARMA's ACF and PACF correlograms via cross-correlation functions. Note that research was not focused on developing the best fit model for each technique, rather we wish to compare in general terms the different approaches to handle system detector in order to estimate missing data. Therefore, we did not perform exhaustive model selection procedures when developing certain models (Time Series, STARMA).

4.5 Evaluation Criteria

The performance of the models employed to estimate missing values of traffic data shall be evaluated by several realistic and informative measures. The following evaluation criteria should expose the relevance of these models to researchers as well as practitioners in the traffic-engineering domain.

4.5.1 Statistical Measure: Mean Absolute Percentage Error

A useful measure of accuracy of an estimation model is to express error as a percentage of deviation from predicted versus actual values. We employ the mean absolute percentage error (MAPE) statistic to compare the fits of the implemented estimation models. This is a suitable statistic to both researcher and practitioner since fluctuations in traffic parameters, such as volume, is usually expressed in terms of percentage change from one time interval to another, vis-à-vis a precise numerical value.

4.5.2 Practical Measure: Traffic Responsive MAPE (V+KO)

Most signal control systems implemented in the United States are based on First Generation or 1.5 Generation Control systems. One mode of operation is to Traffic Responsive (TRSP), which selects the signal-timing plan best suited to current traffic patterns. The benefit of TRSP is that signal-timing plans are automatically implemented based upon traffic demand. However, TRSP mode will not operate if too many detectors fail.

First we calculate the V+KO values for each time-of-day using observed values, and then compare against V+KO values using estimated values. The objective is to minimize the MAPE between the actual and estimated values to substantiate that predicted values are a good substitute for actual data when detectors fail.

5.0 C-STARMA Model

5.1 Model Background

The Times Series and STARMA models used time series data from system detectors to estimate missing data for a given location. Recall that the missing data point is observed for the interval at time *t*. These models utilized previous data from this detector up to this interval (e.g., *t-1*, *t-2*, *t-3*, etc.) depending upon the autoregressive and/or moving average components of the time series models. The STARMA models used previous data (*t-1*, *t-2*, etc.) from spatially related data sources (i.e., neighboring detectors).

These models do not take advantage of contemporaneous data. That is, even if data is missing at time interval *t* at the specified detector, data for time interval *t* may be available at neighboring detectors. We would like to take advantage of as much relevant data as possible to derive estimations of the missing data at the detector of interest. A new model was developed to estimate missing data at time interval *t* for any particular detector by regressing upon the following input parameters: (1) times series data at the detector-of-interest up to time interval *t* (time-series models), (2) data at time interval *t* from spatially correlated neighboring detectors (spatial regression), and (3) time series data from neighboring detectors up to time interval *t* (STARMA models). Using the STARMA model as a foundation, we extend that method by factoring in data at the time interval *t* from spatially correlated neighboring data sources. The C-STARMA model includes input parameters to factor in contemporaneous data. This new model can be viewed as a combination of the STARMA model and the spatial regression model implemented earlier in this research. The notation for the C-STARMA model is as follows:

Notation:

$z_i(t)$	Observation of the random variable $Z_i(t)$
t	Time index ($t = 1, .., T$)
N	Number of fixed sites in space ($l = 1, 2, .., N$)
$L^{(l)}$	Spatial lag operator of spatial order l

The C-STARMA model is expressed as follows:

$$z_i(t) = \beta_0 + \sum_{l=1}^{N} \beta_l z_l(t) + \sum_{k=1}^{p} \sum_{l=0}^{\lambda_k} \phi_{kl} W^{(l)} z_l(t-k) + \varepsilon_i(t) + \sum_{k=1}^{q} \sum_{l=0}^{m_k} \theta_{kl} W^{(l)} \varepsilon_l(t-k)$$

where,

β_0	is the regression constant
β_l	is the regression parameter of the l^{th} term (*note: temporal lag, $k = 0$*)
p	is the autoregressive order,
q	is the moving average order,
λ_k	is the spatial order of the k^{th} autoregressive term,
m_k	is the spatial order of the k^{th} moving-average term
ϕ_{kl}	is the autoregressive parameter at temporal lag k and spatial lag l,
θ_{kl}	is the moving-average parameter at temporal lag k and spatial lag l,
$W^{(l)}$	is the N x N matrix of weights for spatial order l, and
$\varepsilon(t)$	is the random normally distributed error vector at time t

5.2 C-STARMA Model Methodology

This section presents the C-STARMA model building procedure, which is based upon the development of the STARMA model. However, we propose extensions to the STARMA model by introducing enhancements to the model components as well as using a more robust method for parameter estimation. Specifically, factors are introduced into the model to account for contemporaneous data, and stepwise regression practices are used for parameter estimation in case of linear models.

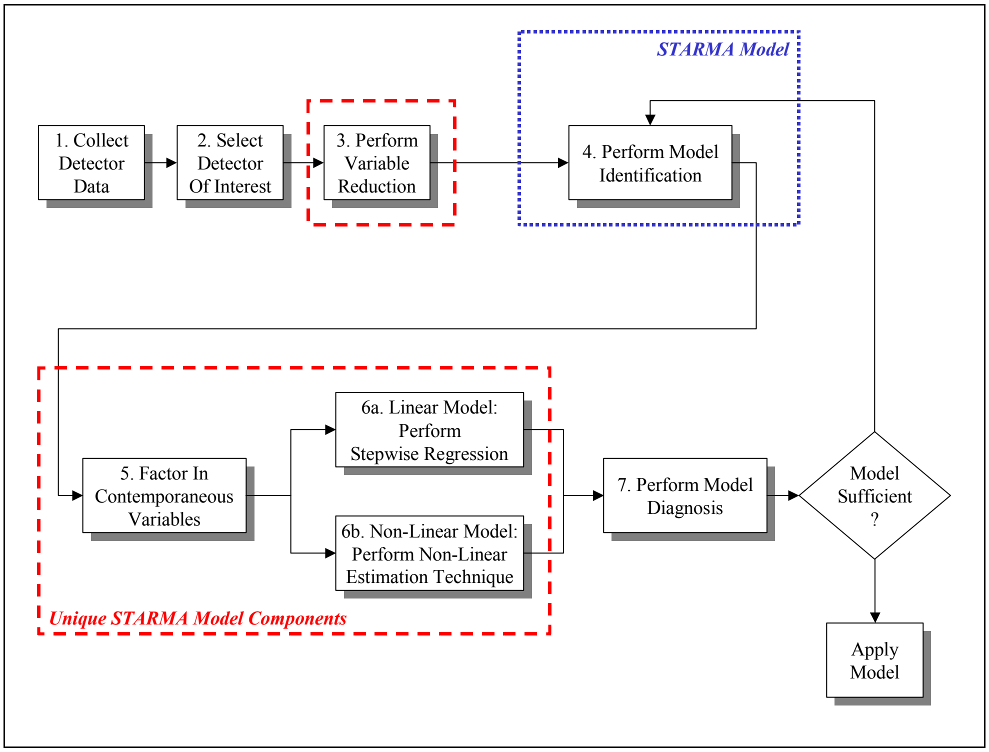

STARMA Model
1. Collect Detector Data
2. Select Detector Of Interest
3. Perform Variable Reduction
4. Perform Model Identification
5. Factor In Contemporaneous Variables
6a. Linear Model: Perform Stepwise Regression
6b. Non-Linear Model: Perform Non-Linear Estimation Technique
Unique STARMA Model Components
7. Perform Model Diagnosis
Model Sufficient ?
Apply Model

Figure 5-1: C-STARMA Model Building Procedure

5.2.1 Collect Detector Data

The initial step in the model building procedure is to collect the data pertinent to the system under study. In our domain of traffic signal systems, system detector data are archived in the Smart Travel Lab's databases. Data should be collected for the location (detector) of interest as well as from other detectors within the arterial network. Detector data should be observed at a regular and corresponding time interval for all the detectors contributing to the data set. This model exploits the spatial correlation between the detector of interest and neighboring detectors and, therefore, data should be collected from those system detectors within the same corridor or network. This research collected volume and occupancy data from detectors within the Reston Area Network, which is the arterial network of system detectors in the Reston, Virginia area.

5.2.2 Select Detector of Interest

The detector of interest is the site or location for which we shall build the C-STARMA model to estimate traffic data. For practical purposes, this is the detector that exhibits missing data for which the analyst desires to impute substitute values. This model building procedure produces a model specific to the detector of interest. Therefore, models would need to be derived for each detector in the network that is deemed critical to the operations of the traffic signal control systems.

5.2.3 Perform Variable Reduction

After the data collection process, the data set may comprise of a large number of detectors in addition to the location of interest. The purpose of this step is to reduce the list of detectors, which serve as inputs to the C-STARMA model. One important benefit of this step is that it exposes the spatial correlation between the detector of interest and its neighboring detectors. This step is analogous to the feature selection or variable selection process in regression models. As such, we apply the same feature selection approach as in the regression model by using the Mallow's C_p criterion. The C_p statistic is defined as follows:

$$C_p = \frac{RSS_p}{\hat{\sigma}^2} + 2p - n$$

where,

n	is the number of observations
p	is the number of variables in the regression
RSS_p	is the residual sum of squares using p variables
$\hat{\sigma}^2$	is an independent estimate of the error

The residual variance from the full model is used as the estimate of $\hat{\sigma}^2$. If the model is satisfactory, C_p will be approximately equal to p. We then select the p variables to proceed with building the C-STARMA model.

5.2.4 Model Identification

The process of identifying the C-STARMA model type is the same as that of the STARMA. The C-STARMA is based upon the STARMA model so the same procedure is applied to identify the autoregressive and/or moving average components of the new model. This process was detailed in Section 3.2.4. The result of this step is the identification of the STARMA components of the C-STARMA model.

5.2.5 Factor In Contemporaneous Variables

This new step in the model building procedure is one of the extensions introduced by the C-STARMA model. The C-STARMA model factors in additional variables to exploit contemporaneous data. These variables account for the data at the latest time interval, t, for each of the input data sources (detectors). The new variables are parameterized by the β coefficients, which are derived by the regression modeling approach. As an example, consider the C-STAR(2_{10}) model:

$$z(t) = \beta_0 + \beta_{00} z(t) + \beta_{01} z(t) + \phi_{10} z(t-1) + \phi_{11} W^{(1)} z(t-1) + \phi_{20} z(t-2) + \varepsilon(t)$$

In general linear model form $\mathbf{Y} = \mathbf{XB} + \varepsilon$, this model for $t = 1, 2, .., T$ can be written as follows:

$$\begin{bmatrix} z(1) \\ z(2) \\ z(3) \\ \vdots \\ z(T) \end{bmatrix} = \beta_0 + \begin{bmatrix} z(1) & z(1) & 0 & 0 & 0 \\ z(2) & Z(2) & z(1) & W^{(1)}z(1) & 0 \\ z(3) & z(3) & z(2) & W^{(1)}z(2) & z(1) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ z(T) & z(T) & z(T-1) & W^{(1)}z(T-1) & z(T-2) \end{bmatrix} \begin{bmatrix} \beta_{00} \\ \beta_{01} \\ \phi_{10} \\ \phi_{11} \\ \phi_{20} \end{bmatrix} + \begin{bmatrix} \varepsilon(1) \\ \varepsilon(2) \\ \varepsilon(3) \\ \vdots \\ \varepsilon(T) \end{bmatrix}$$

The zero vectors were substituted for the unobserved z vectors, for those times before the system was under observation. It should be noted that due to the time series nature of the STAR model, the linear regression assumptions about the independent or regressor variables do not hold. Specifically, the X matrix is stochastic rather than deterministic in repeated samples. Also, the residuals are not normally independently distributed with mean zero and constant variance. The linear regression results will be used here with a priori knowledge that they are only approximate.

5.2.6 Parameter estimation

After a candidate model form has been selected, the next phase of the modeling procedure is to estimate the β, ϕ, and θ parameters. We employ different techniques to approximate the parameter coefficients depending upon whether the C-STARMA model is linear or non-linear. The linear model features only parameters for the regression and autoregressive components, or namely the β and ϕ coefficients. The non-linear form of the C-STARMA model includes the moving average components, which are characterized by the θ coefficients. As with the time series and STARMA models, techniques based on standard linear regression theory are suitable for estimation of the linear model parameters. Due to the non-linear form of the C-STARMA model, non-linear optimization techniques must be employed to estimate the associated parameters.

Another modification introduced by the C-STARMA model is the use of stepwise regression for parameter selection and inclusion in the final model. Recall that the STARMA model procedure includes all detectors as input variables in its model, and then eliminates variables in a backward manner that are deemed statistically insignificant. The C-STARMA procedure implements stepwise regression to iteratively build the model by adding and eliminating variables according to the statistical entry and removal criteria.

5.2.7 Perform Model Diagnosis

After a candidate model has been selected and its parameters estimated the model must undergo evaluation to determine whether the model adequately represents the data. The first phase is to perform an analysis of the residuals to verify that they are approximately normally and independently distributed with zero mean and σ^2 variance. The second phase is to verify the statistical significance of the estimated parameters. In addition, we employ the *Adjusted-*R^2 criterion to evaluate linear models. If the model is judged to fit the available data sufficiently well, then it can be used to impute missing traffic data. However, if the model is insufficient, the analyst must repeat the model-building steps from the model identification stage.

6.0 RESULTS & ANALYSIS

This chapter presents the results of the imputation models for each of the three detector scenarios, as well as the extended network data case. The primary metric used to evaluate the models was the Mean Absolute Percentage Error (MAPE). The MAPE metric was evaluated for estimations of volume (vehicles per hour), occupancy (percentage of hour), and V+KO. As previously stated, a lower MAPE score typically indicates a stronger model to impute missing data. In addition, the error distributions specify the percentage and range of under- and over-estimations of missing data. Ideally, stronger models should indicate a normal distribution of residuals.

6.1 Results of Models Using Upstream Detectors Only: Detector 2001

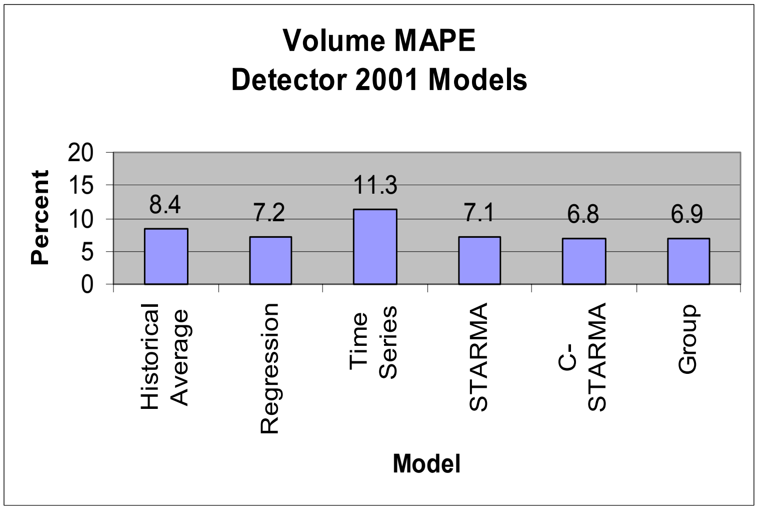
This traffic scenario features the use of only upstream detectors to estimate traffic data at the specified location. Table 6-1 presents the MAPE performance measure for each of the models at estimating volume, occupancy, and V+KO. All models performed well in terms of MAPE values that may be operationally neglible during non-peak periods. For example, the regression model's volume MAPE of 7.2% may not be as significant for periods of 100 vehicles per hour as compared to peak periods where 1000+ vehicles per hour are observed. The C-STARMA model produced the best estimates for volume data, but only marginally came in second to the pure Regression model for estimating occupancy. The Group model produced the best MAPE outcome for the V+KO estimates. As expected, these particular models performed well because they factored in the spatial correlation between the detector of interest and its neighboring detectors.

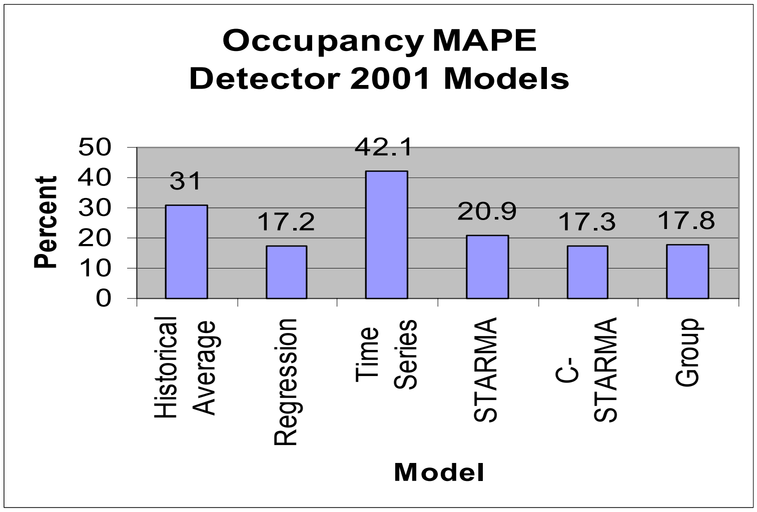
Note that the MAPE results for occupancy models were artificially inflated due to the scale and narrower range of this particular data type. One should be careful to not directly compare the volume and V+KO MAPE numbers to those for occupancy.

Table 6-1: Mean Absolute Percentage Error (MAPE) - Detector 2001 Scenario

Detector 2001 Mean Absolute Percentage Error (%)			
MODEL	**MAPE (Volume)**	**MAPE (Occupancy)**	**MAPE (V+KO)**
Historical Average	8.4	31.0	11.5
Regression	7.2	**17.2**	7.4
Time Series	11.3	42.1	14.6
STARMA	7.1	20.9	8.8
C-STARMA	**6.8**	17.3	7.4
Group	6.9	17.8	**7.3**

The following figures graphically depict the MAPE results presented in Table 6-1. We observe that the models that take advantage of the spatial relationships (e.g., Regression, STARMA, C-STARMA, and Group models) between the detectors generally performed better than the models that relied solely upon univariate data (Historical Average and Time Series models).





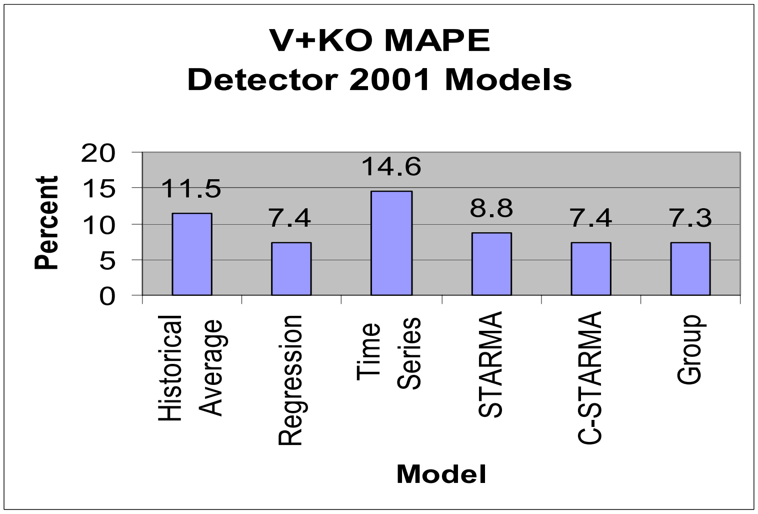


Figure 6-1: MAPE Results - Detector 2001

Table 6-2 presents the distribution of the models' residuals for volume imputation in the detector 2001 scenario. Overall, all the models performed very well by producing residuals that were normally distributed with mean zero and variance $\hat{\sigma}^2$. Each of the models produced estimates that fell within the +/- 15% error range, which imply generally accurate estimations. These results support the strength of the models in addition to their MAPE values. The C-STARMA model was the most precise model by producing the largest number of estimations that fell within the +/- 5% range.

Table 6-2: Error Distributions for Volume - Detector 2001

Detector 2001 Model Error Distribution - Volume							
MODEL	**< -25%**	**-25% to -15%**	**-15% to -5%**	**-5% to 5%**	**5% to 15%**	**15% to 25%**	**> 25%**
Historical Average	0.5	6.0	29.3	37.2	20.4	4.6	2.1
Regression	0.2	3.9	23.0	42.1	24.6	5.6	0.7
Time Series	1.8	8.4	21.6	31.8	20.7	9.0	6.8
STARMA	0.0	2.9	22.4	42.8	25.9	5.1	0.9
C-STARMA	0.0	2.0	26.0	**47.0**	21.0	4.0	1.0
Group	0.0	1.0	16.0	46.0	30.0	6.0	1.0

Figure 6-2 graphically depicts the distribution of the models' residuals. The C-STARMA model's precision is validated by producing the majority of the estimates that fell within the +/- 15% error.

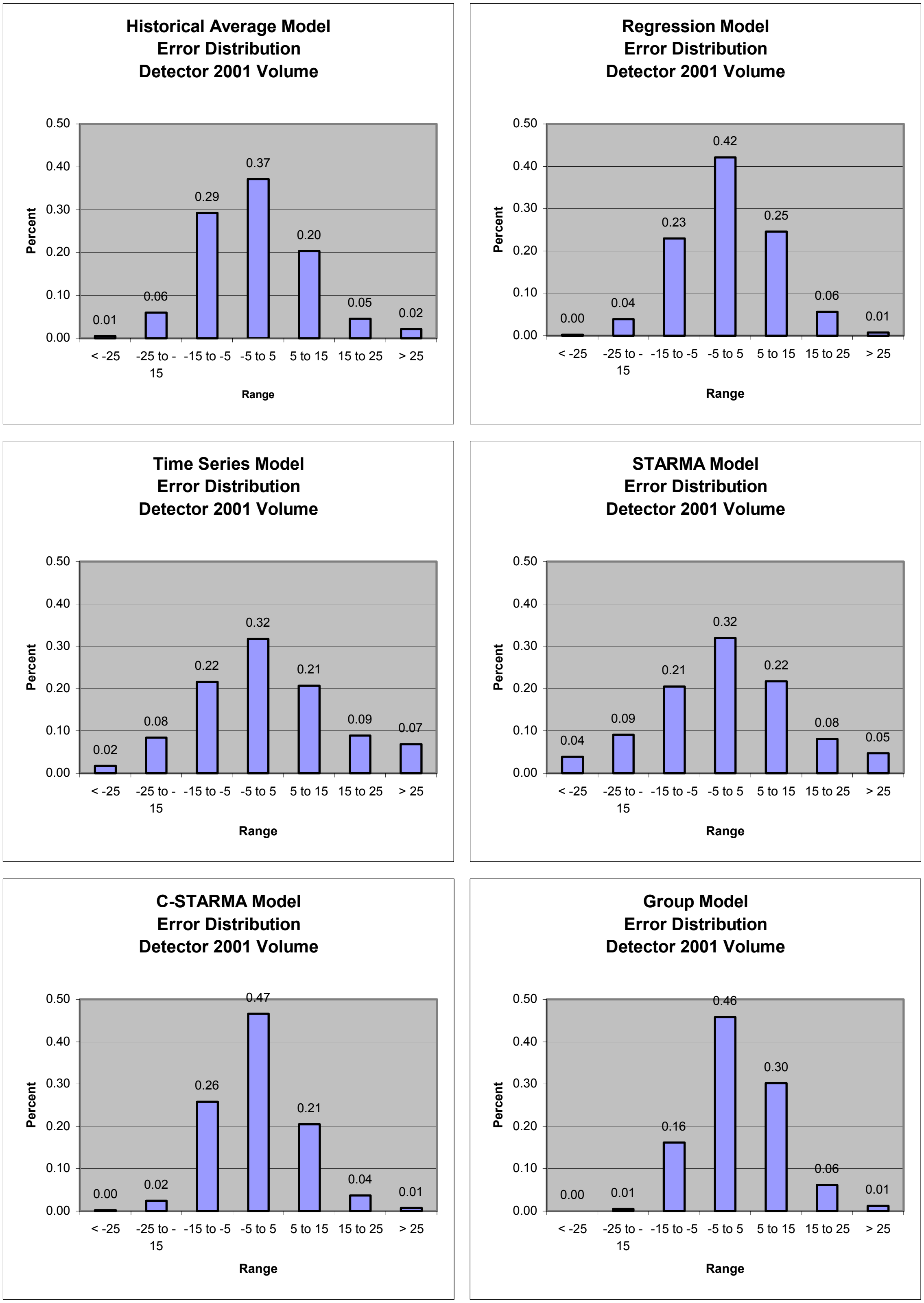


Figure 6-2: Volume Error Distribution - Detector 2001 Models

Table 6-3 presents the distribution of the models' residuals for occupancy imputation in the detector 2001 scenario. The residual distributions for the occupancy estimations were more widespread as compared to those for volume. Recall that volume measures can reach to several thousand vehicles per hour, whereas occupancy is rated on a percentage scale. Though the error distributions for occupancy are more distributed over the error ranges, this does not necessarily indicate that the models performed poorly. The rational for the widely distributed errors can be attributed to the lower numeral scale for occupancy data. For example, if the actual occupancy value for a particular time interval was 5 compared to an estimation of 6, the resulting error is 20. This error percentage of 20 in occupancy is not comparable to a 20 error in volume or V+KO. A normalization process should be performed to properly compare the error percentages across the three metrics.

Table 6-3: Error Distributions for Occupancy - Detector 2001

Detector 2001 Model Error Distribution - Occupancy							
MODEL	**< -25%**	**-25% to -15%**	**-15% to -5%**	**-5% to 5%**	**5% to 15%**	**15% to 25%**	**> 25%**
Historical Average	26.14	11.58	8.07	16.14	14.21	9.82	14.04
Regression	8.60	11.75	15.61	21.75	20.18	7.89	14.21
Time Series	22.46	10.53	10.70	10.18	9.12	8.07	28.95
STARMA	27.2	8.8	8.8	9.0	6.1	7.2	33.0
C-STARMA	7.00	13.0	14.0	**24.0**	18.0	10.0	14.0
Group	6.00	13.0	16.0	22.0	17.0	9.00	16.0

Figure 6-3 graphically depicts the residual distributions presented the table above. Though residuals were evenly distributed across all the models, the C-STARMA produced the largest number of estimations within the +/-5%. Interestingly, the STARMA and Time Series models produced the least reliable estimates of occupancy, which was also reflected in their residual distributions.

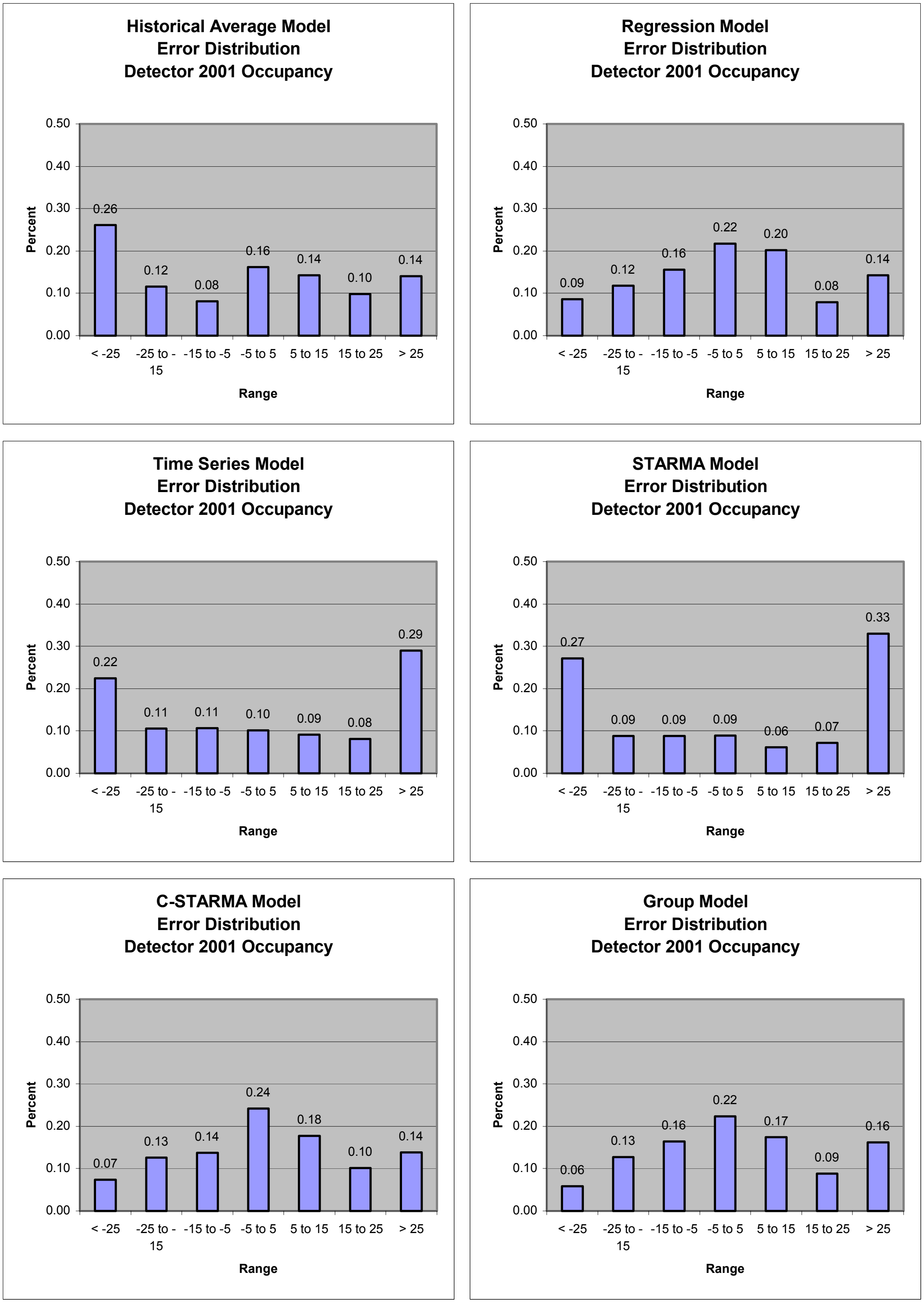


Figure 6-3: Occupancy Error Distribution - Detector 2001 Models

The V+KO metric was included to demonstrate the practical validity and application of these missing data estimation models. 1.5 Generation Systems and beyond use the V+KO method for traffic adaptive signal control systems. In addition, this metric supports the strength of the models at estimating both volume and occupancy. Table 6-4 presents the error distributions for the estimating V+KO in the detector 2001 scenario.

Table 6-4: Error Distributions for V+KO - Detector 2001

Detector 2001 Error Distribution - V+KO (where K=20)							
MODEL	**< -25%**	**-25% to -15%**	**-15% to -5%**	**-5% to 5%**	**5% to 15%**	**15% to 25%**	**> 25%**
Historical Average	1.6	10.0	29.0	27.7	20.2	6.5	5.1
Regression	0.2	5.1	21.6	42.1	24.6	5.4	1.1
Time Series	4.0	9.0	20.5	24.6	19.8	9.8	12.3
STARMA	0.7	3.6	20.4	42.4	25.1	6.0	1.6
C-STARMA	0.0	5.0	24.0	**45.0**	21.0	5.0	1.0
Group	0.0	1.0	18.0	**45.0**	28.0	8.0	1.0

We observe that the models that used multiple predictor variables performed better than models that solely relied on univariate data. These models produced results that predominantly fell within the +/-15% error. In addition, the stronger models were the ones that took advantage of the spatial correlation among the neighboring detectors and the point of interest. Figure 6-4 graphically depicts the residual distributions for models imputing the V+KO data. The precision of the C-STARMA and Group models were again validated due to the large percentage of estimations within the +/-15% range.

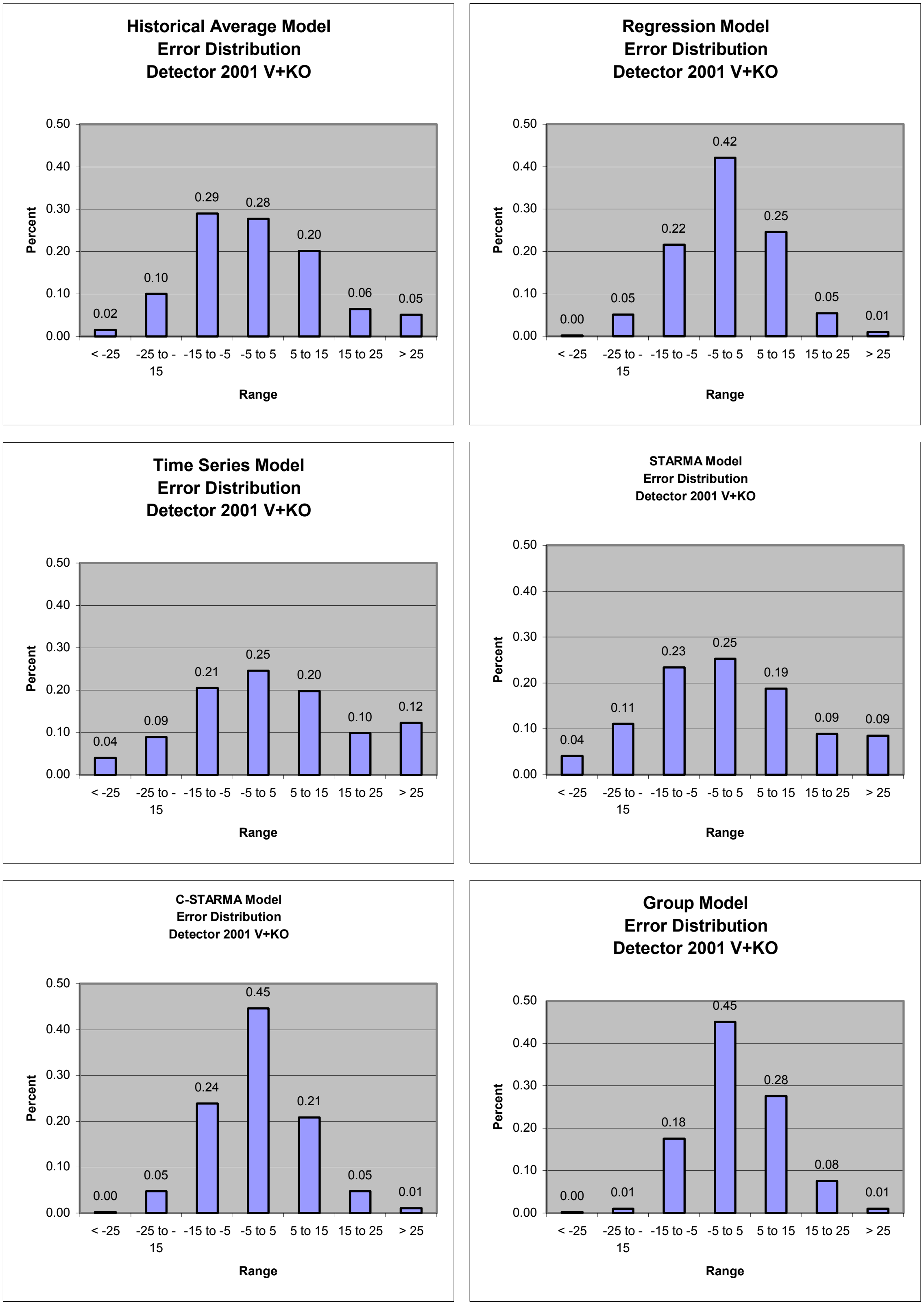


Figure 6-4: V+KO Error Distribution - Detector 2001 Models

6.2 Results of Models Using Both Upstream and Downstream Detectors: Detector 2027

This traffic scenario features the use of both upstream and downstream detectors to estimate traffic data at the specified location. Table 6-5 presents the MAPE results of the models imputing volume, occupancy, and V+KO for system detector 2027. Figure 6-5 graphically depicts the MAPE results from the table. The STARMA model produced the best results for volume and V+KO imputation, while the Group produced the best MAPE result for occupancy estimates. Other models that produced comparatively low MAPE numbers included the Regression and C-STARMA models. Again, this supports the hypothesis that the better models take advantage of the spatial relationship among neighboring detectors.

Table 6-5: Mean Absolute Percentage Error (MAPE) - Detector 2027 Scenario

Detector 2027 Mean Absolute Percentage Error (%)			
MODEL	**MAPE (Volume)**	**MAPE (Occupancy)**	**MAPE (V+KO)**
Historical Average	8.8	15.7	9.4
Regression	5.7	20.3	6.3
Time Series	18.8	21.4	20.7
STARMA	**5.5**	12.7	**6.1**
C-STARMA	7.1	12.1	7.2
Group	9.3	**11.7**	9.0

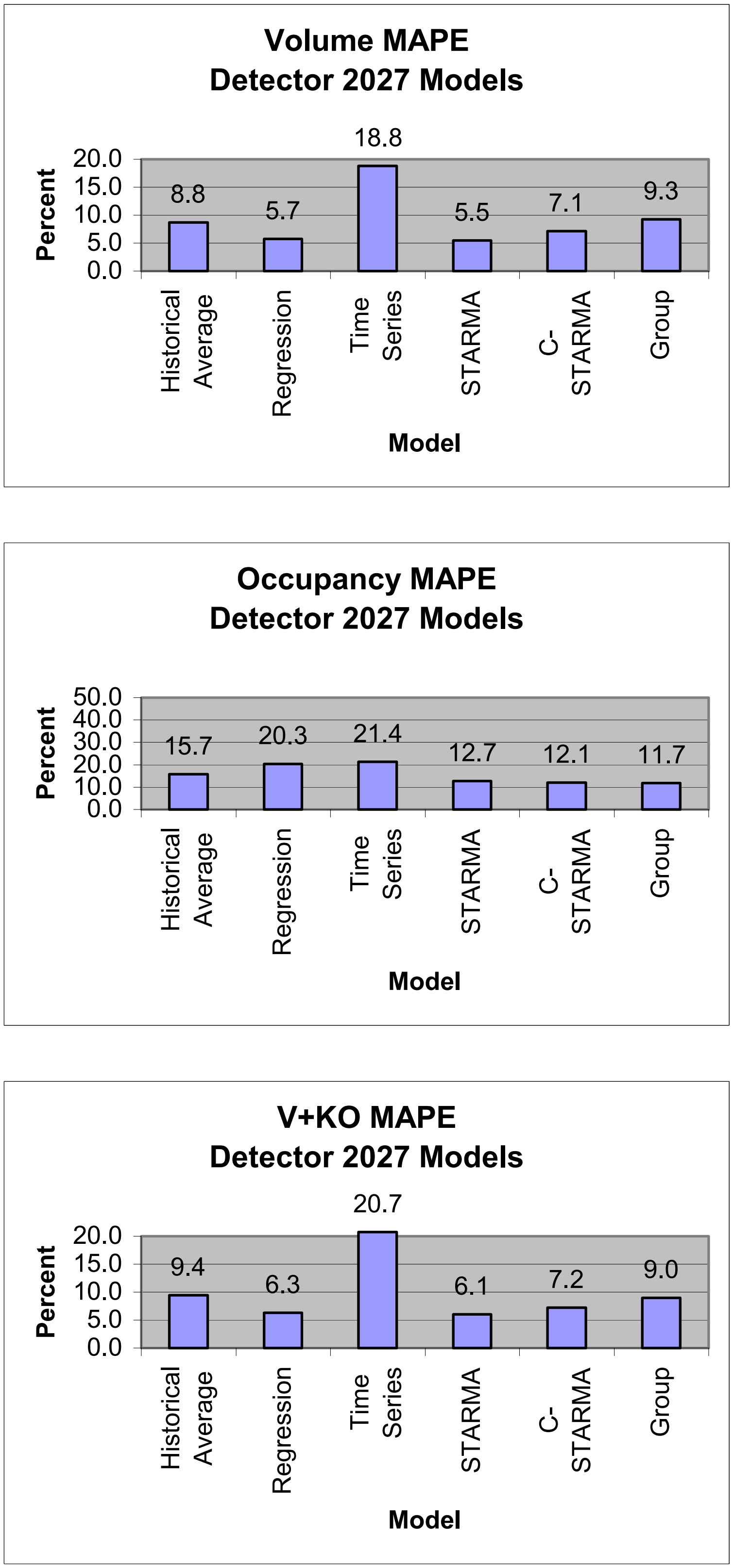


Figure 6-5: MAPE Results - Detector 2027

Table 6-6 presents the distribution of the models' residuals for volume imputation in the detector 2027 scenario. Figure 6-6 illustrates the residual distributions. Each of the models' residual was generally normally distributed, although the Time Series model's residuals were more widespread. The STARMA model produced the largest percentage of estimates within the +/-5% range, and was closely followed by the Regression, C-STARMA, and Group models. In addition, the majority of the results from these models fell within the +/-15% range. The multi-variate models that took into consideration the spatial characteristics of the data outperformed the univariate models.

Table 6-6: Error Distributions for Volume - Detector 2027

Detector 2027 Model Error Distribution - Volume							
MODEL	**< -25%**	**-25% to -15%**	**-15% to -5%**	**-5% to 5%**	**5% to 15%**	**15% to 25%**	**> 25%**
Historical Average	1.1	5.8	26.3	37.4	20.5	5.8	3.2
Regression	0.0	1.9	20.1	57.9	16.8	1.9	1.5
Time Series	3.9	7.2	23.7	25.4	20.4	10.2	9.3
STARMA	0.0	1.1	17.7	**59.0**	18.8	2.4	0.9
C-STARMA	0.0	3.0	21.0	49.0	19.0	4.0	3.0
Group	1.0	3.0	17.0	56.0	17.0	1.0	5.0

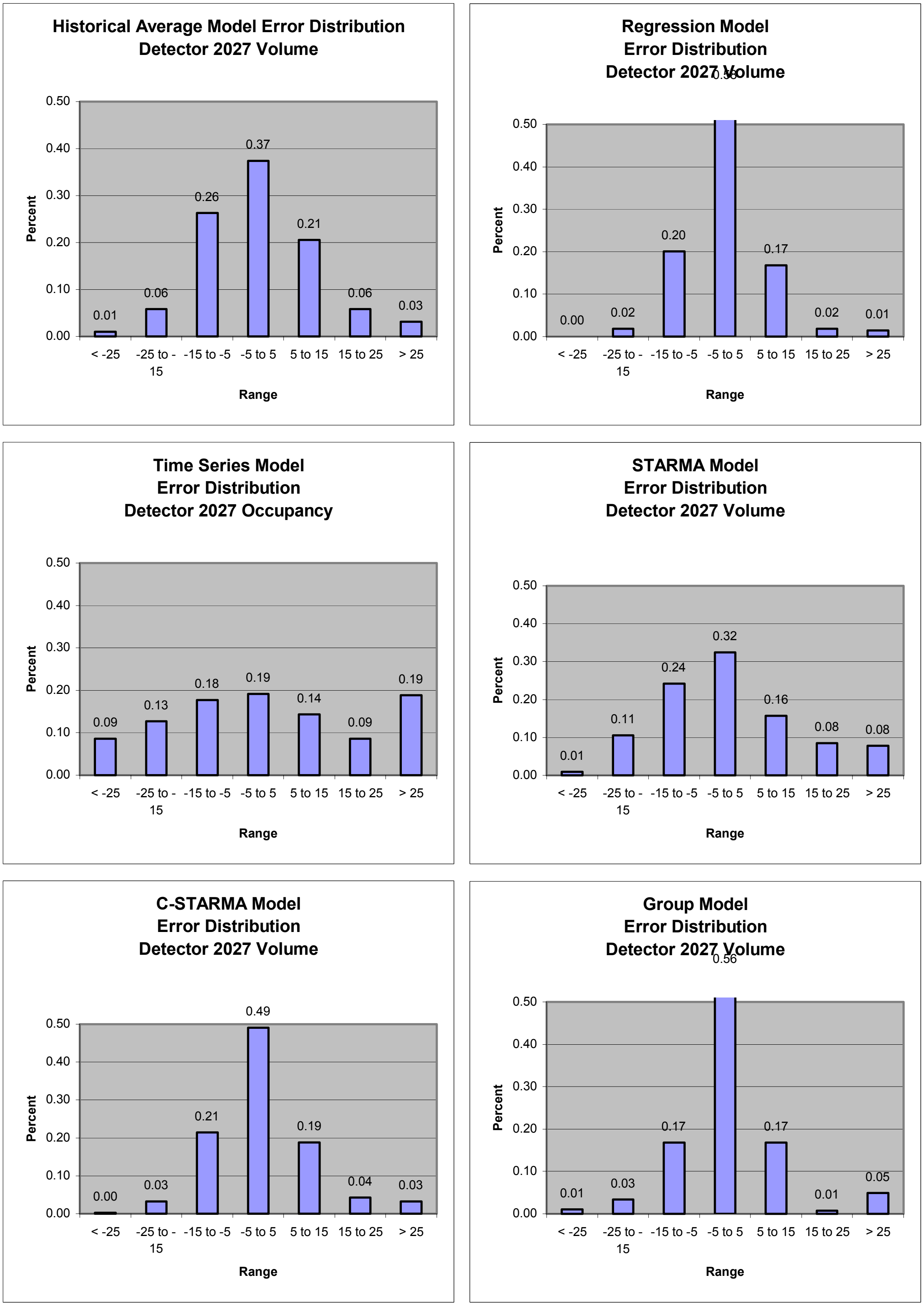


Figure 6-6: Volume Error Distribution - Detector 2027 Models

Table 6-7 and Figure 6-7 present the residual distributions for occupancy imputation in the detector 2027 scenario. The residuals are distributed across the ranges for most of the models, except for the C-STARMA and Group model, which are relatively normally distribution. The Group model produced the largest percentage of estimates within the +/-5% range, and is closely followed by the C-STARMA model.

Table 6-7: Error Distributions for Occupancy - Detector 2027

Detector 2027 Model Error Distribution - Occupancy							
MODEL	**< -25%**	**-25% to -15%**	**-15% to -5%**	**-5% to 5%**	**5% to 15%**	**15% to 25%**	**> 25%**
Historical Average	6.3	12.7	26.3	24.6	9.9	14.2	6.5
Regression	4.7	11.3	19.1	12.8	20.6	9.8	21.7
Time Series	8.6	12.7	17.7	19.2	14.3	8.6	18.8
STARMA	3.0	11.1	23.6	24.7	20.2	8.7	8.7
C-STARMA	2.0	6.0	15.0	32.0	22.0	14.0	9.0
Group	7.0	6.0	18.0	**33.0**	24.0	10.0	2.0

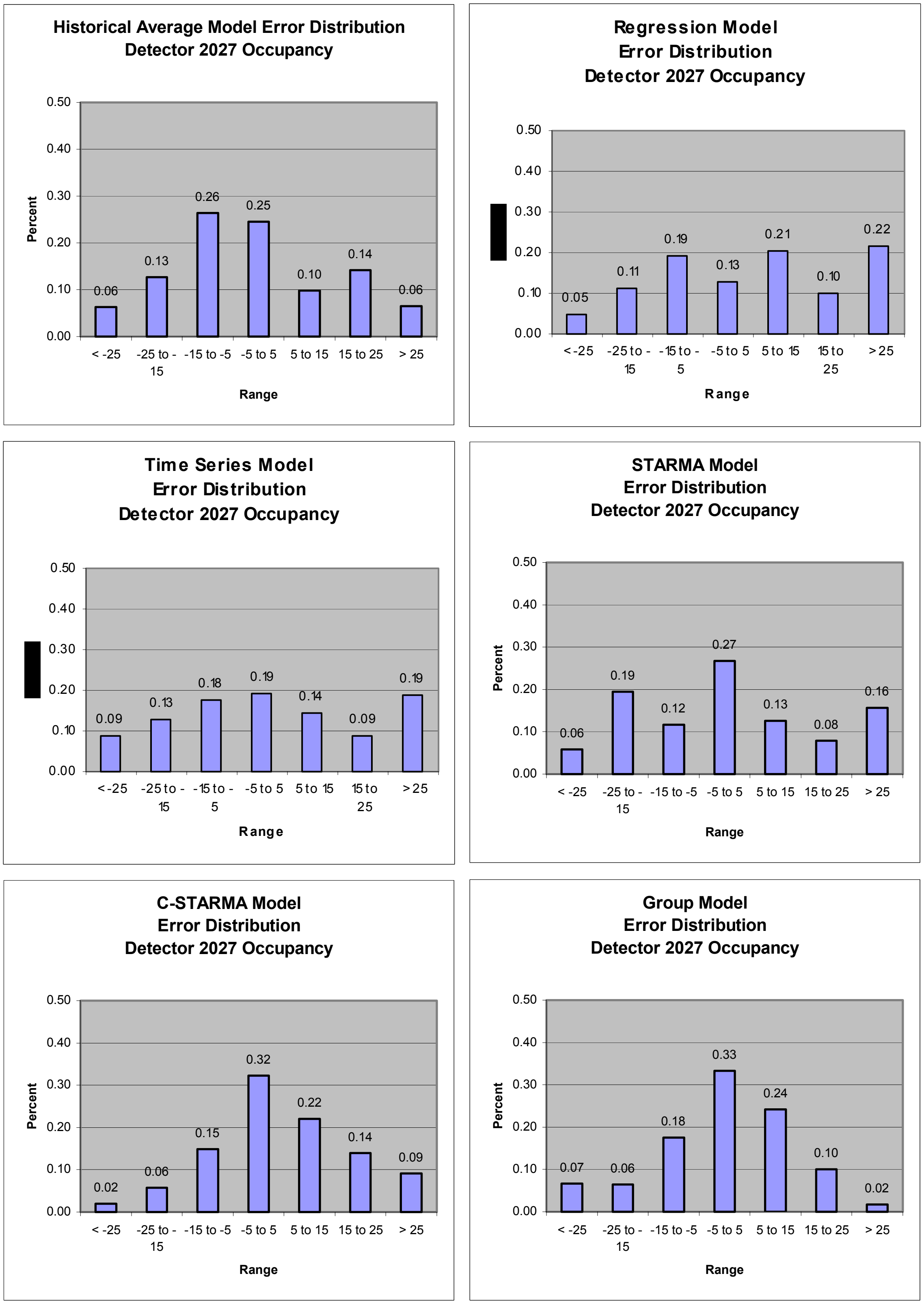


Figure 6-7: Occupancy Error Distribution - Detector 2027 Models

Table 6-8 presents the distribution of the models' residuals for V+KO imputation in the detector 2027 scenario. Again, we observe that multi-variate models produced residuals that are ideally normally distributed and outperformed the univariate models. The Group model produced the largest percentage of estimations within +/-5% range, and closely followed by the C-STARMA and Regression models. Figure 6-8 illustrates the residual distributions of these models. The models that emphasized the spatial nature of the detector data (Regression, C-STARMA, and Group models) indicated significant precision in imputation of V+KO data.

Table 6-8: Error Distributions for V+KO - Detector 2027

	Detector 2027 Error Distribution - V+KO						
MODEL	**< -25%**	**-25% to -15%**	**-15% to -5%**	**-5% to 5%**	**5% to 15%**	**15% to 25%**	**> 25%**
Historical Average	1.1	6.8	25.8	36.0	19.1	7.5	3.7
Regression	0.0	0.6	21.2	50.4	23.1	2.6	2.2
Time Series	4.2	6.5	24.7	24.7	17.9	10.9	11.1
STARMA	1.2	9.5	25.3	31.8	16.5	8.1	7.7
C-STARMA	0.0	2.0	18.0	52.0	19.0	5.0	3.0
Group	5.0	1.0	17.0	**55.0**	17.0	9.0	1.0

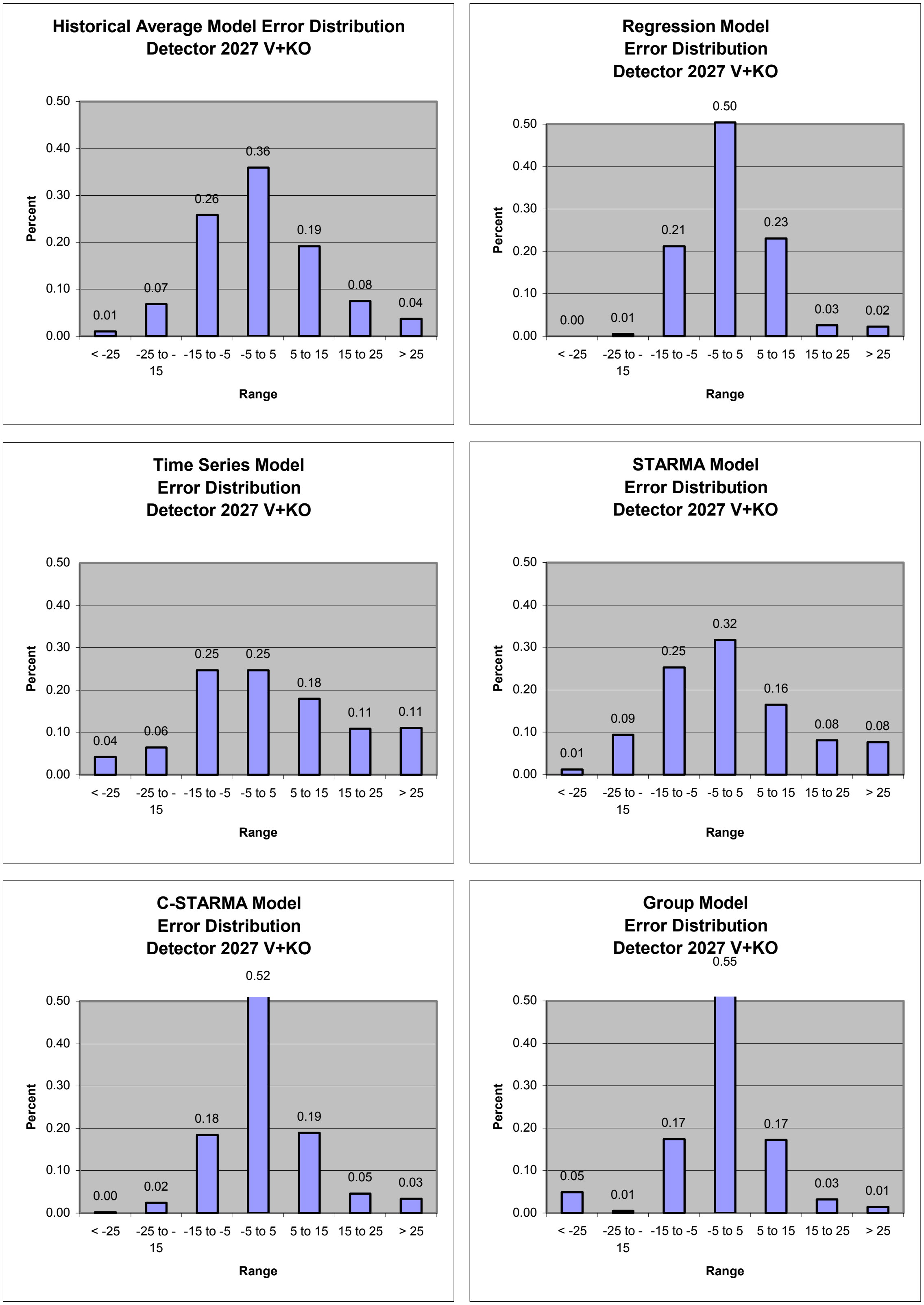


Figure 6-8: V+KO Error Distribution - Detector 2027 Models

6.3 Results of Models Using Downstream Detectors Only: Detector 2037

This traffic scenario features the use of only downstream detectors to estimate traffic data at the specified location. Table 6-9 presents the MAPE performance measure for each of the models at estimating volume, occupancy, and V+KO for the detector 2037 scenario. As with the other scenarios, we observe that the multivariate, spatial models outperformed the univariate models. The C-STARMA produced the lowest MAPE numbers for volume and V+KO estimations, while the Group model performed the best at estimating occupancy data. Figure 6-9 illustrates the MAPE results for each of the models.

Table 6-9: Mean Absolute Percentage Error (MAPE) - Detector 2037 Scenario

Detector 2037 Mean Absolute Percentage Error (%)			
MODEL	**MAPE (Volume)**	**MAPE (Occupancy)**	**MAPE (V+KO)**
Historical Average	8.3	12.3	8.8
Regression	6.7	12.2	9.0
Time Series	12.4	17.3	12.9
STARMA	6.5	10.7	10.8
C-STARMA	**6.3**	10.8	**6.6**
Group	6.4	**10.1**	**6.6**

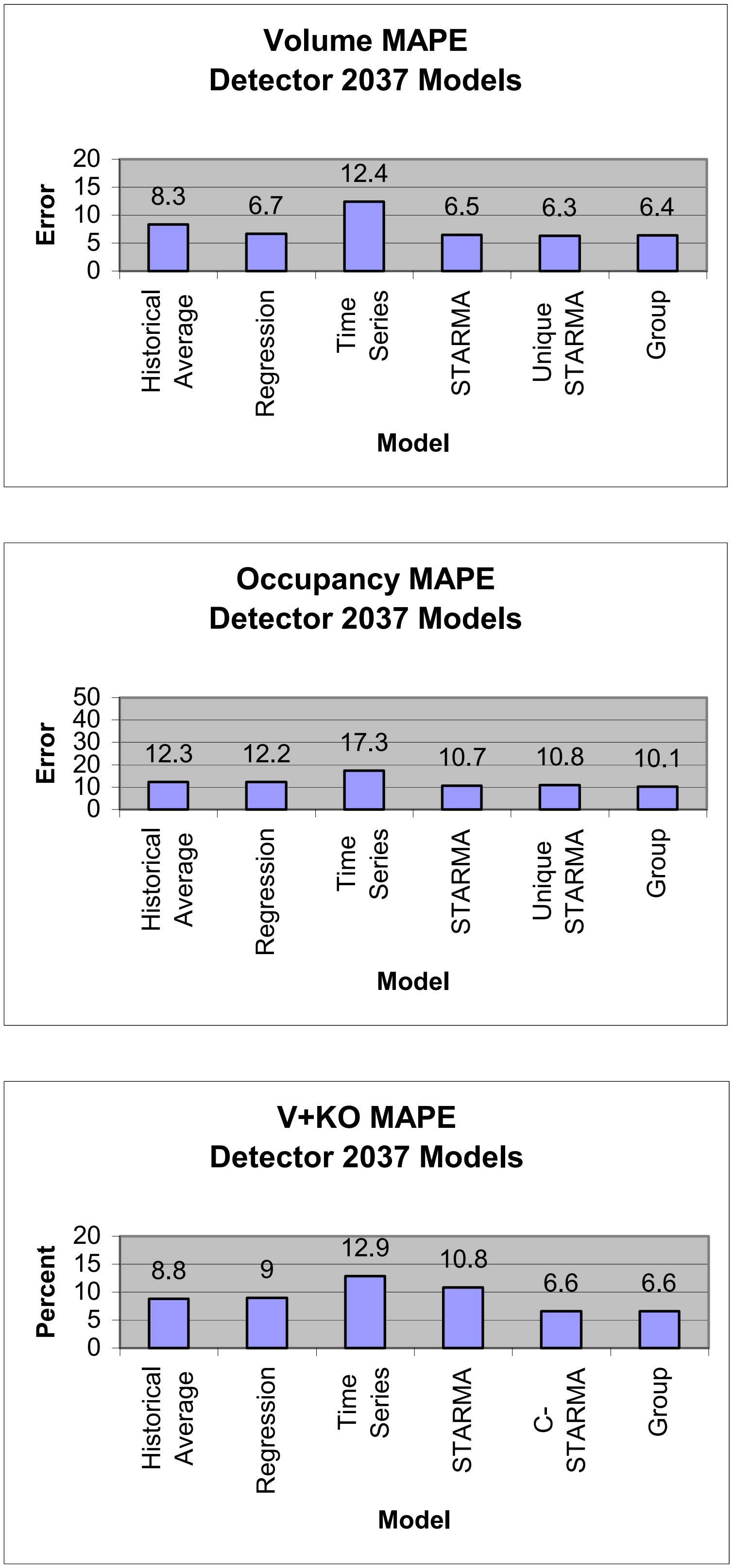


Figure 6-9: MAPE Results - Detector 2037

Table 6-10 presents the distribution of the detector 2037 volume residuals. Figure 6-10 illustrates how each of the models' residuals is normally distributed. We again observe that the multivariate spatial models produced the largest percentages of estimations within the +/-5% range.

Table 6-10: Error Distributions for Volume - Detector 2037

Detector 2037 Model Error Distribution – Volume							
MODEL	**< -25%**	**-25% to -15%**	**-15% to -5%**	**-5% to 5%**	**5% to 15%**	**15% to 25%**	**> 25%**
Historical Average	0.5	6.1	27.4	38.8	19.3	6.5	1.4
Regression	0.2	3.9	24.9	48.8	17.7	4.4	0.2
Time Series	1.8	9.8	23.2	28.6	18.6	9.7	8.8
STARMA	0.2	2.0	22.3	50.6	19.2	5.5	0.2
C-STARMA	0.0	2.0	22.0	52.0	20.0	5.0	0.0
Group	0.0	5.0	18.0	**53.0**	21.0	3.0	0.0

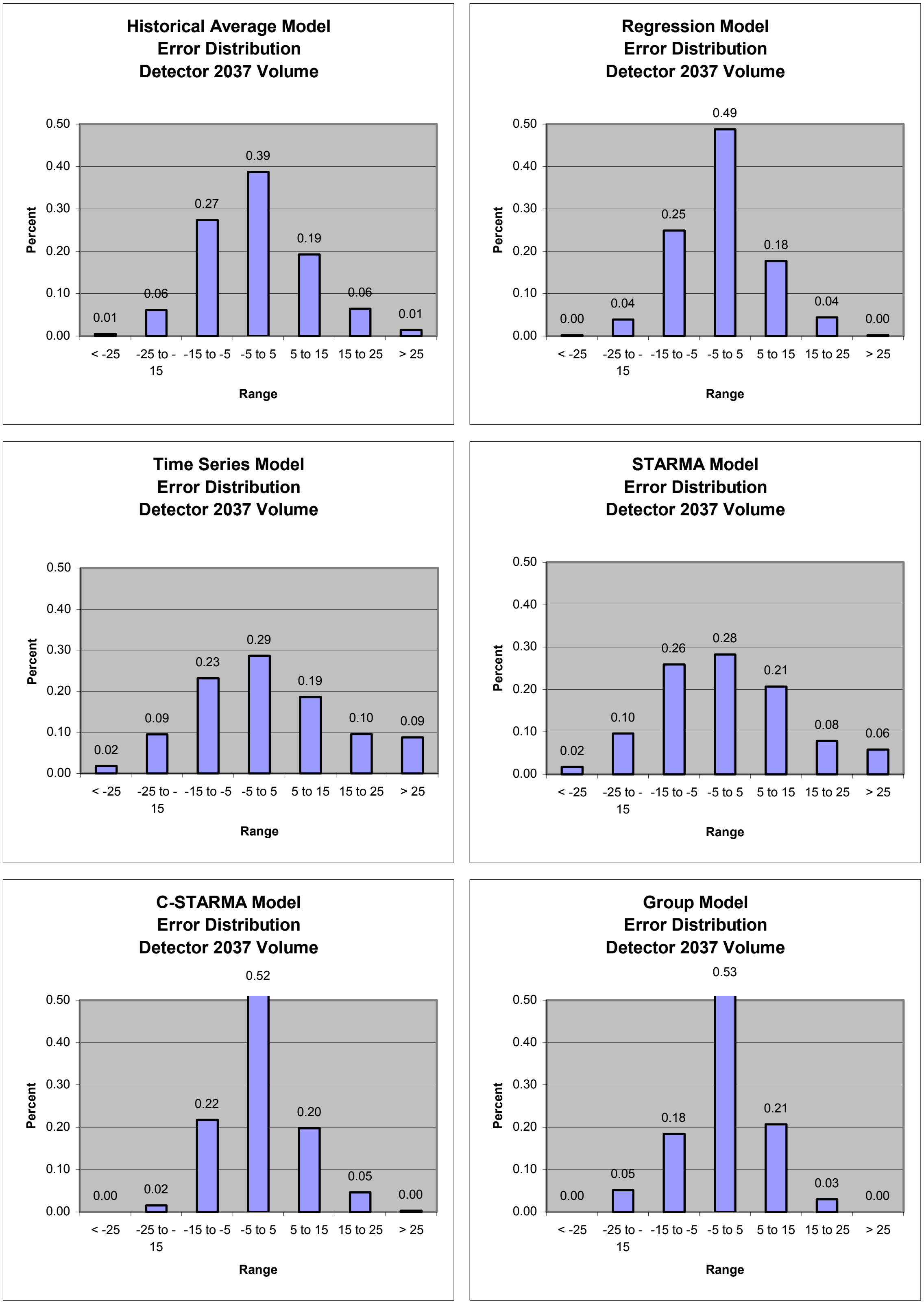


Figure 6-10: Volume Error Distribution - Detector 2037 Models

Table 6-11 and Figure 6-11 present the error distributions for estimating occupancy in the detector 2037 scenario. The residual distribution for this particular model was closer to the ideal normal distribution than those residual found in the detector 2001 and 2027 scenarios. Although residuals were more distributed across the ranges, a larger percentage of these models' errors concentrated in the +/-15% range. The Group model produced the largest number of estimations within the +/-5% range.

Table 6-11: Error Distributions for Occupancy - Detector 2037

Detector 2037 Model Error Distribution – Occupancy							
MODEL	**< -25%**	**-25% to -15%**	**-15% to -5%**	**-5% to 5%**	**5% to 15%**	**15% to 25%**	**> 25%**
Historical Average	4.2	13.2	24.9	27.4	15.7	8.4	6.1
Regression	3.2	13.4	25.7	27.9	14.6	8.0	7.1
Time Series	6.0	12.5	18.8	23.0	13.9	12.3	13.7
STARMA	0.9	7.7	24.5	30.4	20.0	11.3	5.2
C-STARMA	2.0	11.0	26.0	30.0	20.0	7.0	4.0
Group	5.0	8.0	20.0	**35.0**	22.0	8.0	1.0

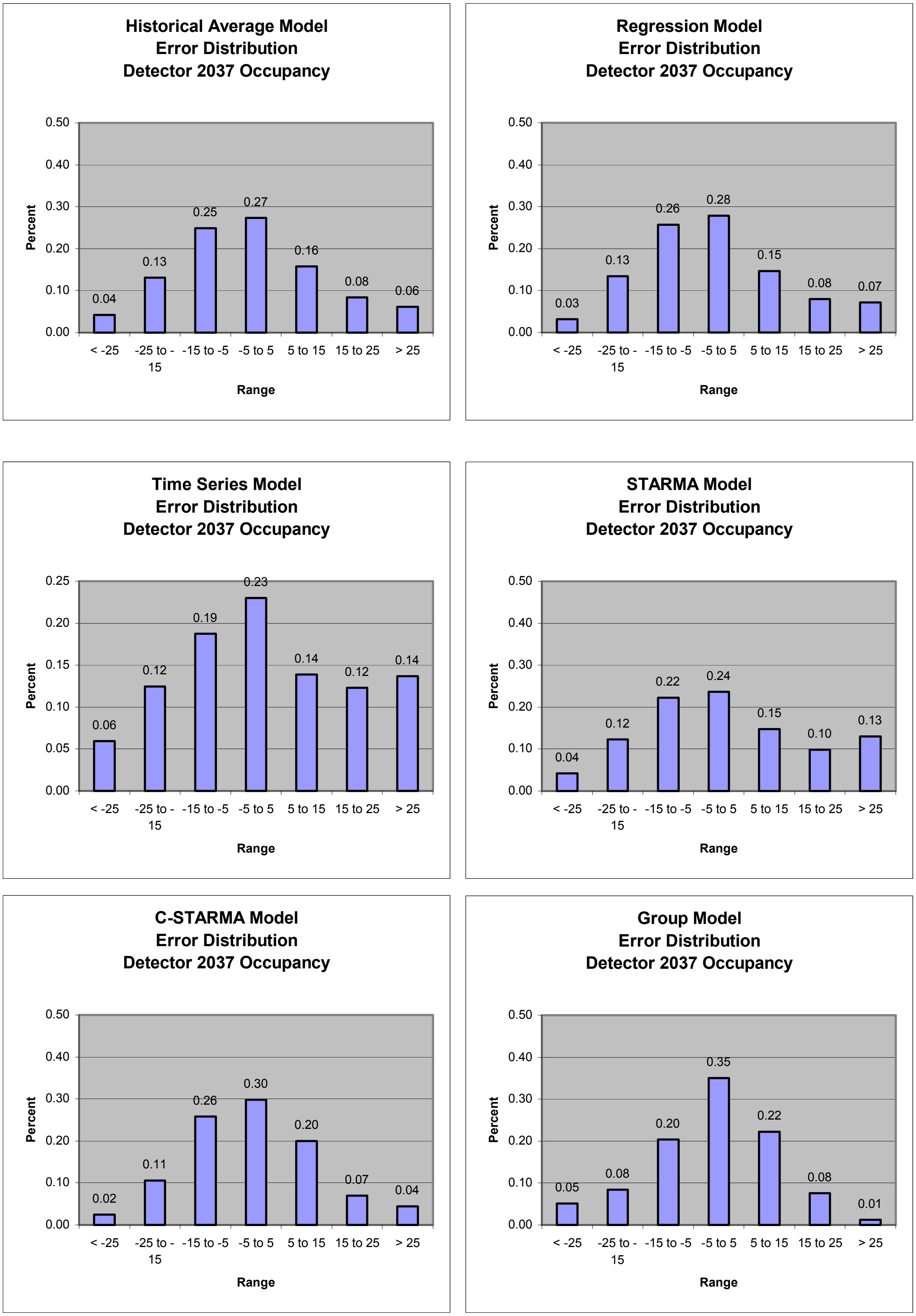


Figure 6-11: Occupancy Error Distributions - Detector 2037 Models

Table 6-12 presents the distribution of the models' residuals for V+KO imputation in the detector 2037 scenario. Again, we observe that multi-variate models produced residuals that are ideally normally distributed and outperformed the univariate models. The C-STARMA and Group models produced the largest percentage of estimations within +/-5% range. Figure 6-12 illustrates the residual distributions of these models. The models that emphasized the spatial nature of the detector data (Regression, STARMA, C-STARMA, and Group models) indicated significant precision in imputation of V+KO data.

Table 6-12: Error Distributions for V+KO - Detector 2037

Detector 2037 Error Distribution - V+KO							
MODEL	**< -25%**	**-25% to -15%**	**-15% to -5%**	**-5% to 5%**	**5% to 15%**	**15% to 25%**	**> 25%**
Historical Average	0.7	7.4	30.2	35.3	17.5	6.8	2.1
Regression	3.0	3.6	27.5	44.4	16.4	4.8	0.4
Time Series	1.9	10.0	23.2	27.0	19.0	9.1	9.8
STARMA	4.4	1.6	22.5	45.4	19.3	6.7	0.2
C-STARMA	0.0	2.0	24.0	**49.0**	19.0	6.0	0.0
Group	0.0	3.0	22.0	**49.0**	20.0	6.0	0.0

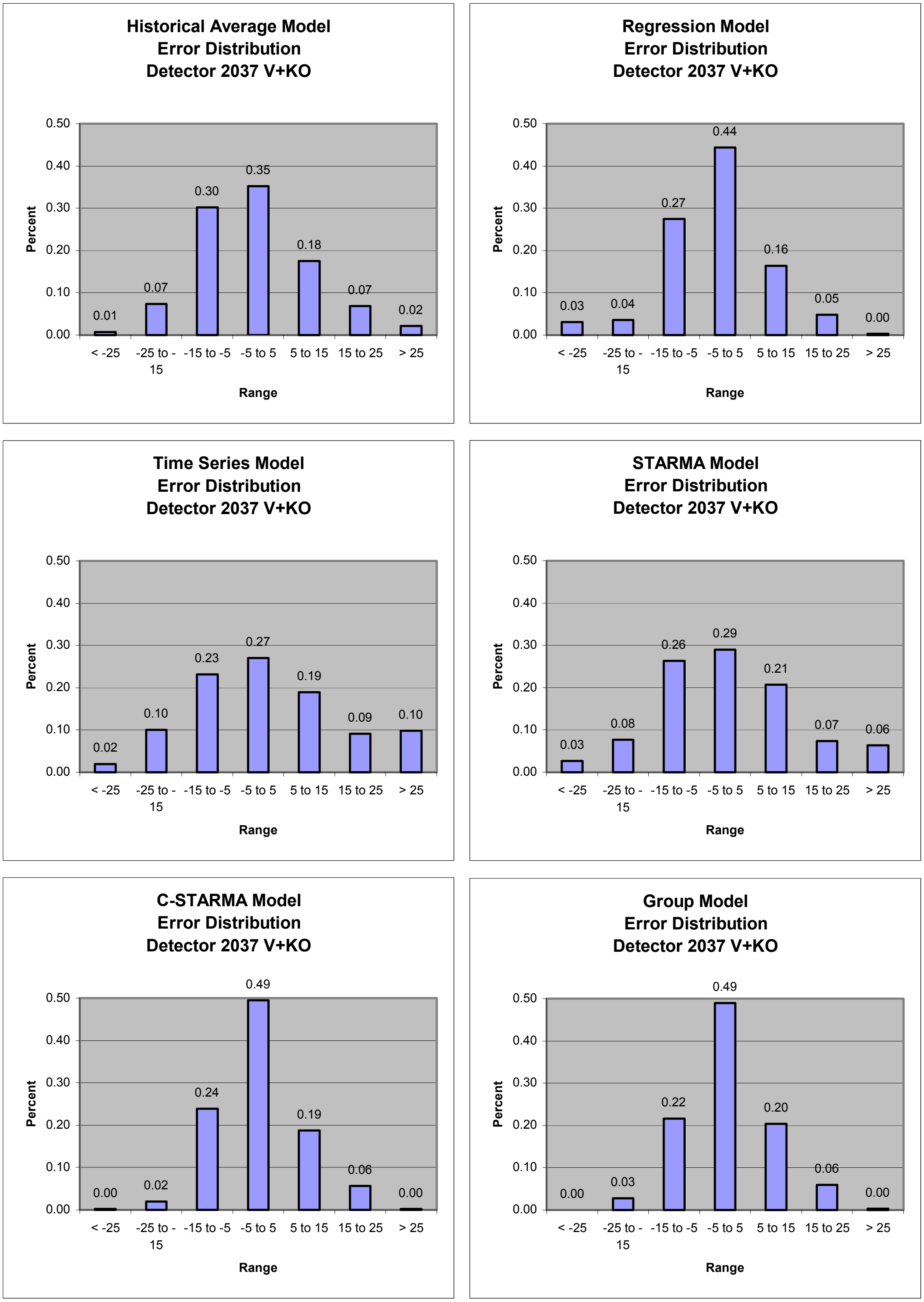


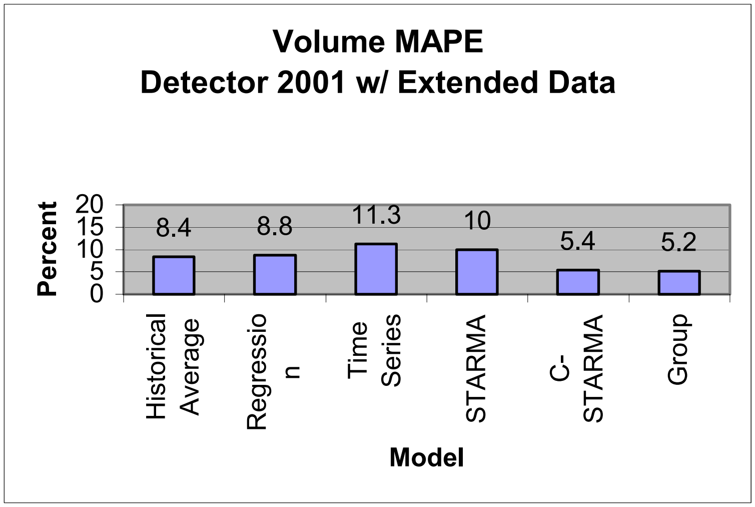
Figure 6-12: V+KO Error Distribution - Detector 2037 Models

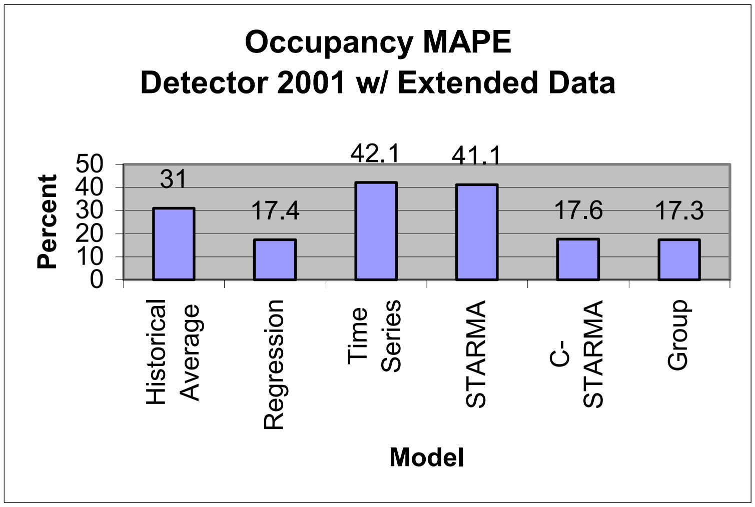
6.4 Results of Models Using Extended Network Data

Table 6-13 and Figure 6-13 present the MAPE results of the detector 2001 models developed using the extended data set from the Reston Area Network. These models imputed traffic data for detector 2001 using detectors throughout the arterial network. Again, we observe that the multivariate spatial models outperformed the univariate models for traffic data imputation. The Group and C-STARMA models produced the best MAPE numbers across the three traffic data elements. This particular scenario validates the utility of the C-STARMA approach over other spatially oriented models, such as the STARMA. Note the comparatively larger value of the occupancy MAPE value of the STARMA model. This may be attributed to the significantly larger number of input variables this model included in its final composition. Recall that the STARMA model includes all input variables and then performs backwards selection to prune its variables. The final STARMA model in this case still included a larger number of input variables than the C-STARMA. Since these variables did not significantly contribute to the strength of the STARMA model, they most likely induced more noise into the model. The C-STARMA occupancy model was more parsimonious in its final variable selection.

Table 6-13: MAPE Results - Detector 2001 Using Extended Data Set

Detector 2001 w/ Extended Network Data - Mean Absolute Percentage Error (%)			
MODEL	**MAPE (Volume)**	**MAPE (Occupancy)**	**MAPE (V+KO)**
Historical Average	8.4	31.0	11.5
Regression	8.8	17.4	8.4
Time Series	11.3	42.1	14.6
STARMA	10.0	41.1	12.5
C-STARMA	5.4	17.6	6.1
Group	**5.2**	**17.3**	**6.0**





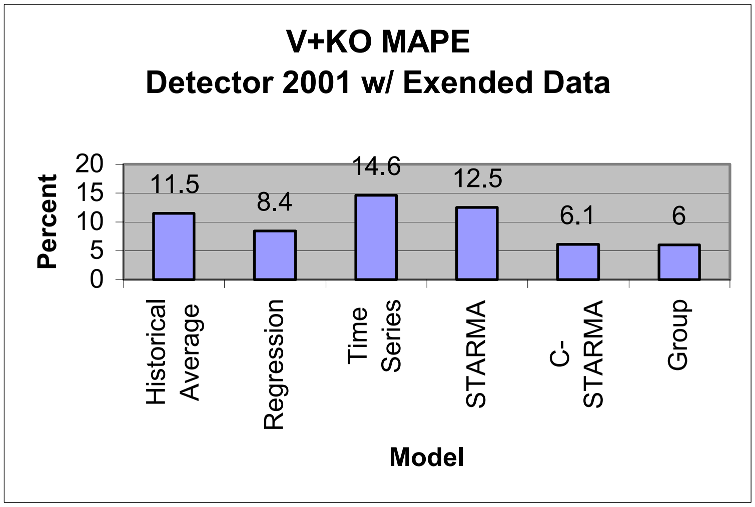


Figure 6-13: MAPE Results - Detector 2001 Models using Extended Data Set

Table 6-14 presents the distribution of the models' residuals for volume imputation in the detector 2001 extended network data scenario. Figure 6-14 illustrates the residual distributions. Each of the models' residual was generally normally distributed, although the Regression model's residuals were skewed negative. The Group and C-STARMA models produced the largest percentage of estimates within the +/-5% range. When exposed to a larger set of data sources, the C-STARMA model's precision was above all other models in imputing volume data. The Group volume model featured a large bias towards the C-STARMA model.

Table 6-14: Volume Error Distribution - Detector 2001 Using Extended Network Data

Detector 2001 w/ Extended Data Set Model Error Distribution - Volume							
MODEL	**< -25%**	**-25% to -15%**	**-15% to -5%**	**-5% to 5%**	**5% to 15%**	**15% to 25%**	**> 25%**
Historical Average	0.5	6.0	29.3	37.2	20.4	4.6	2.1
Regression	4.0	14.0	41.0	37.0	4.0	0.0	0.0
Time Series	1.8	8.4	21.6	31.8	20.7	9.0	6.8
STARMA	4.0	5.0	21.0	38.0	23.0	6.0	4.0
C-STARMA	1.0	2.0	21.0	58.0	18.0	1.0	0.0
Group	0.0	1.0	20.0	**59.0**	19.0	1.0	0.0

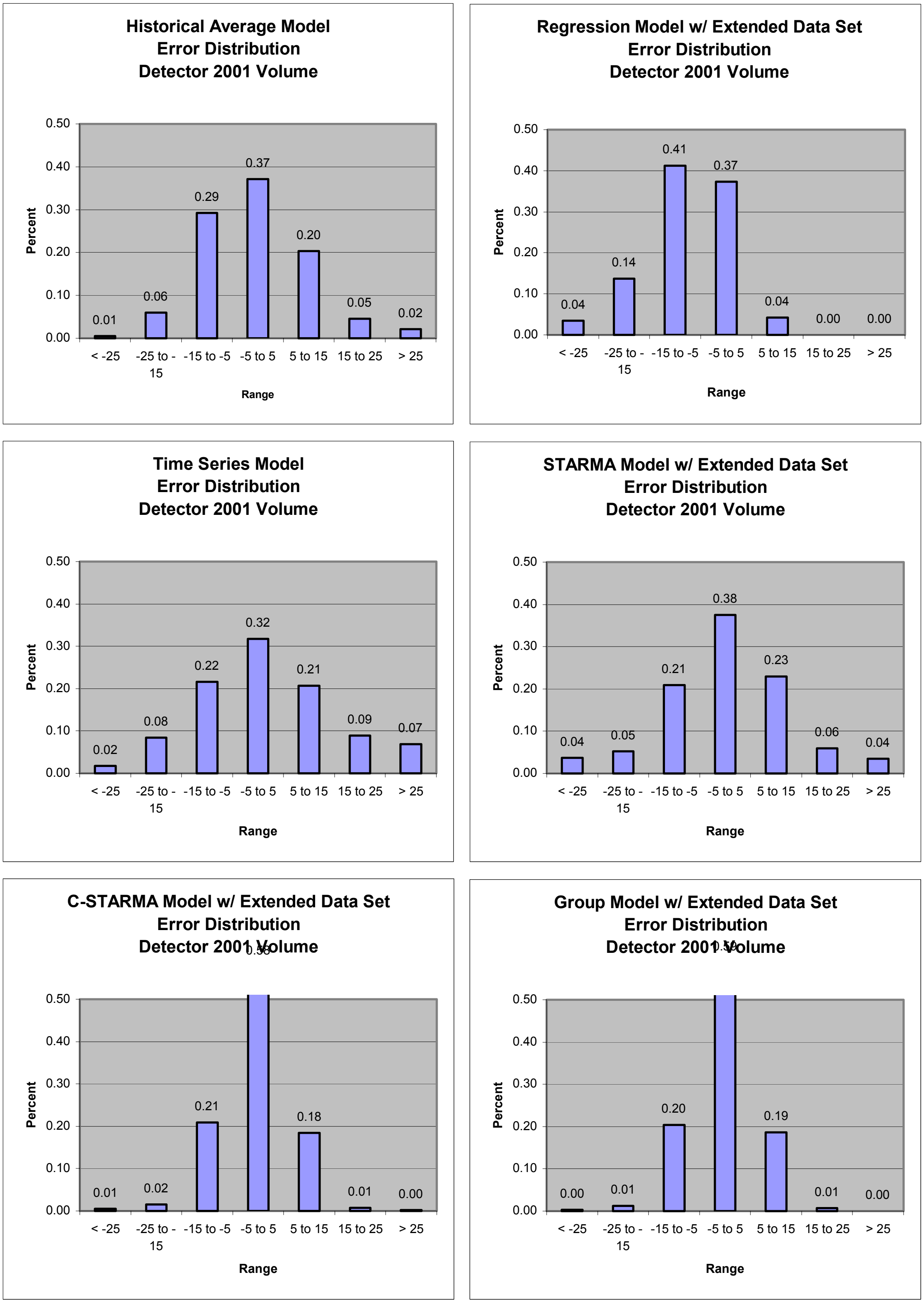


Figure 6-14: Volume Error Distribution - Detector 2001 Models Using Extended Data Set

Table 6-15 and Figure 6-15 present the residual distributions for occupancy imputation in the detector 2001 extended network data scenario. The residuals are distributed across the ranges for most of the models, except for the C-STARMA and Group model, which are relatively normally distribution. The Group model produced the largest percentage of estimates within the +/-5% range, and is closely followed by the C-STARMA model. The Group occupancy model featured a large bias towards the C-STARMA model.

Table 6-15: Occupancy Error Distribution - Detector 2001 Models Using Extended Data Set

Detector 2001 w/ Extended Network Data Model Error Distribution - Occupancy							
MODEL	**< -25%**	**-25% to -15%**	**-15% to -5%**	**-5% to 5%**	**5% to 15%**	**15% to 25%**	**> 25%**
Historical Average	26.1	11.6	8.1	16.1	14.2	9.8	14.0
Regression	13.0	9.0	15.0	24.0	20.0	9.0	11.0
Time Series	22.5	10.5	10.7	10.2	9.1	8.1	29.0
STARMA	29.0	7.0	8.0	11.0	9.0	9.0	27.0
C-STARMA	13.0	9.0	15.0	25.0	19.0	9.0	10.0
Group	15.0	11.0	15.0	**26.0**	17.0	9.0	8.0

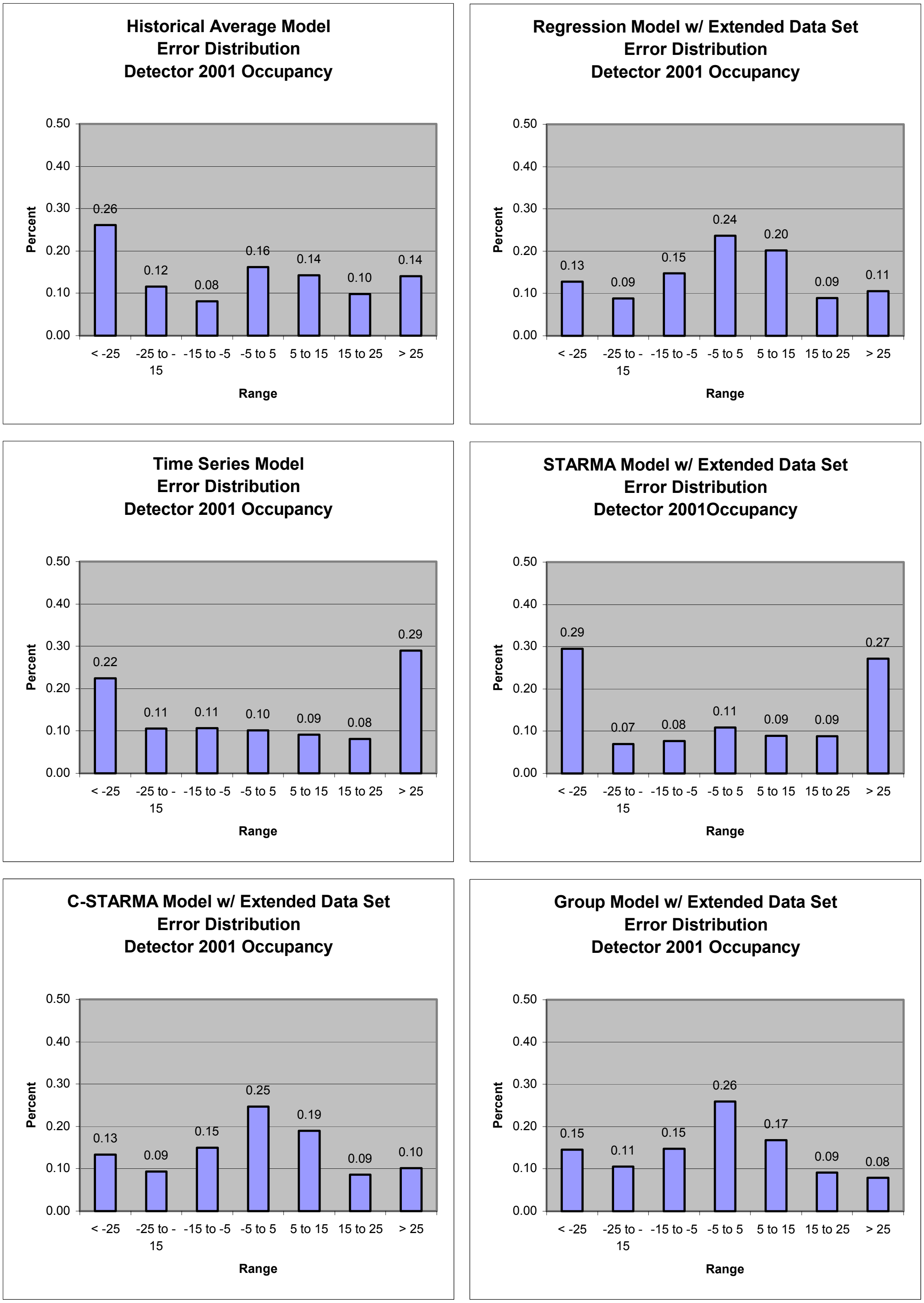


Figure 6-15: Occupancy Error Distribution - Detector 2001 Models Using Extended Data Set

Table 6-16 presents the distribution of the models' residuals for V+KO imputation in the detector 2001 extended network data scenario. All the models' residuals roughly normally distributed, however, it is clear the C-STARMA and Group models outperformed all others by producing the largest percentage of estimations within +/-5% range. Figure 6-16 illustrates the residual distributions of these models. The C-STARMA and Group models indicated significant precision in imputation of V+KO data over the other models.

Table 6-16: V+KO Error Distribution - Detector 2001 Models Using Extended Data Set

Detector 2001 w/ Extended Network Data Error Distribution - V+KO (where K=20)							
MODEL	**< -25%**	**-25% to -15%**	**-15% to -5%**	**-5% to 5%**	**5% to 15%**	**15% to 25%**	**> 25%**
Historical Average	1.6	10.0	29.0	27.7	20.2	6.5	5.1
Regression	0.0	0.0	7.0	39.0	38.0	13.0	3.0
Time Series	4.0	9.0	20.5	24.6	19.8	9.8	12.3
STARMA	6.0	8.0	21.0	30.0	19.0	9.0	6.0
C-STARMA	0.0	2.0	19.0	**53.0**	22.0	2.0	1.0
Group	0.0	1.0	20.0	51.0	24.0	3.0	1.0

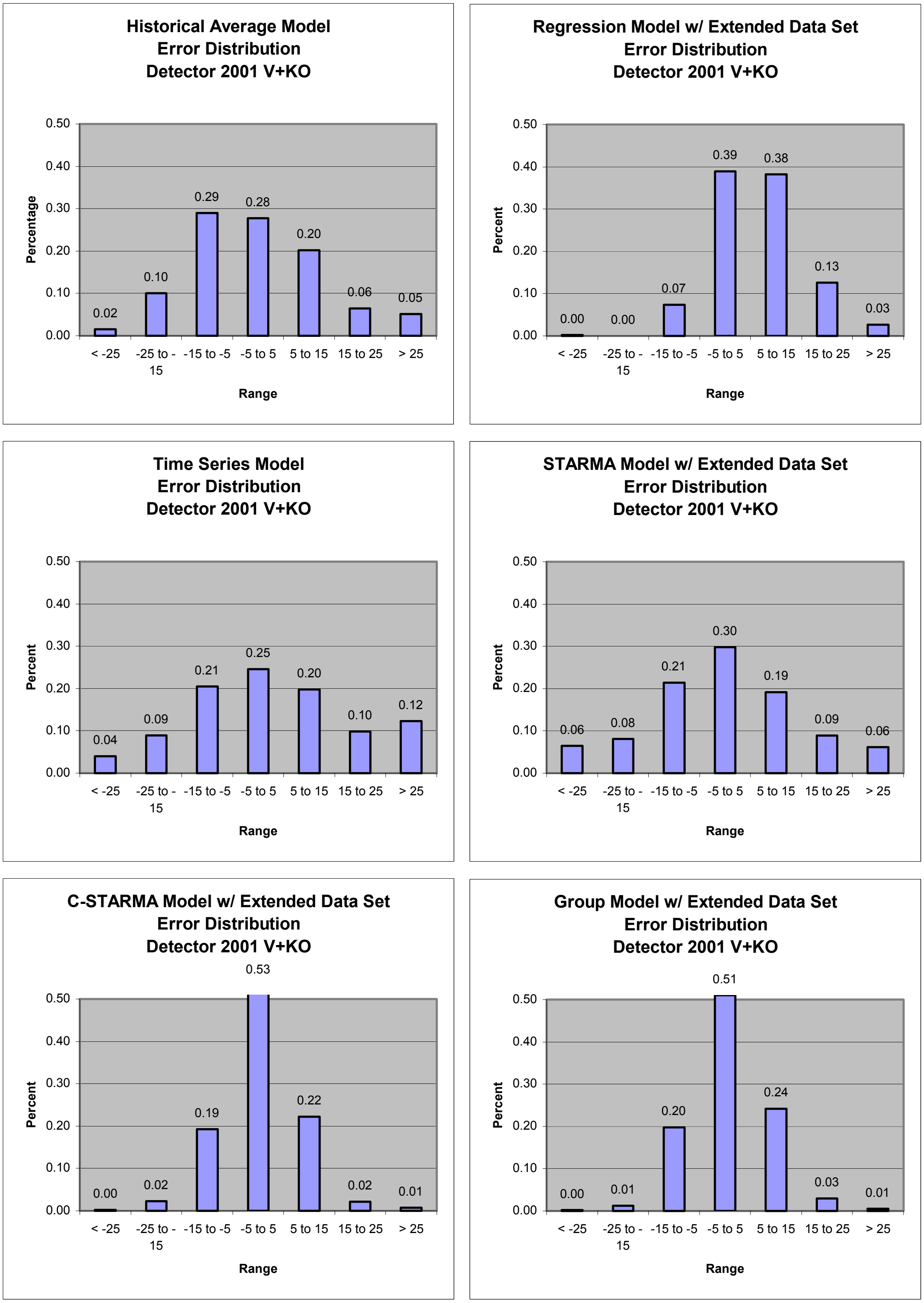


Figure 6-16: V+KO Error Distribution - Detector 2001 Models Using Extended Data Set

Figures 6-17 and 6-18 graphically illustrate which detectors were selected as inputs to the final C-STARMA volume and occupancy models using the extended network data set. Both examples indicate the prudence of the variable selection to create parsimonious models that are accurate and precise.

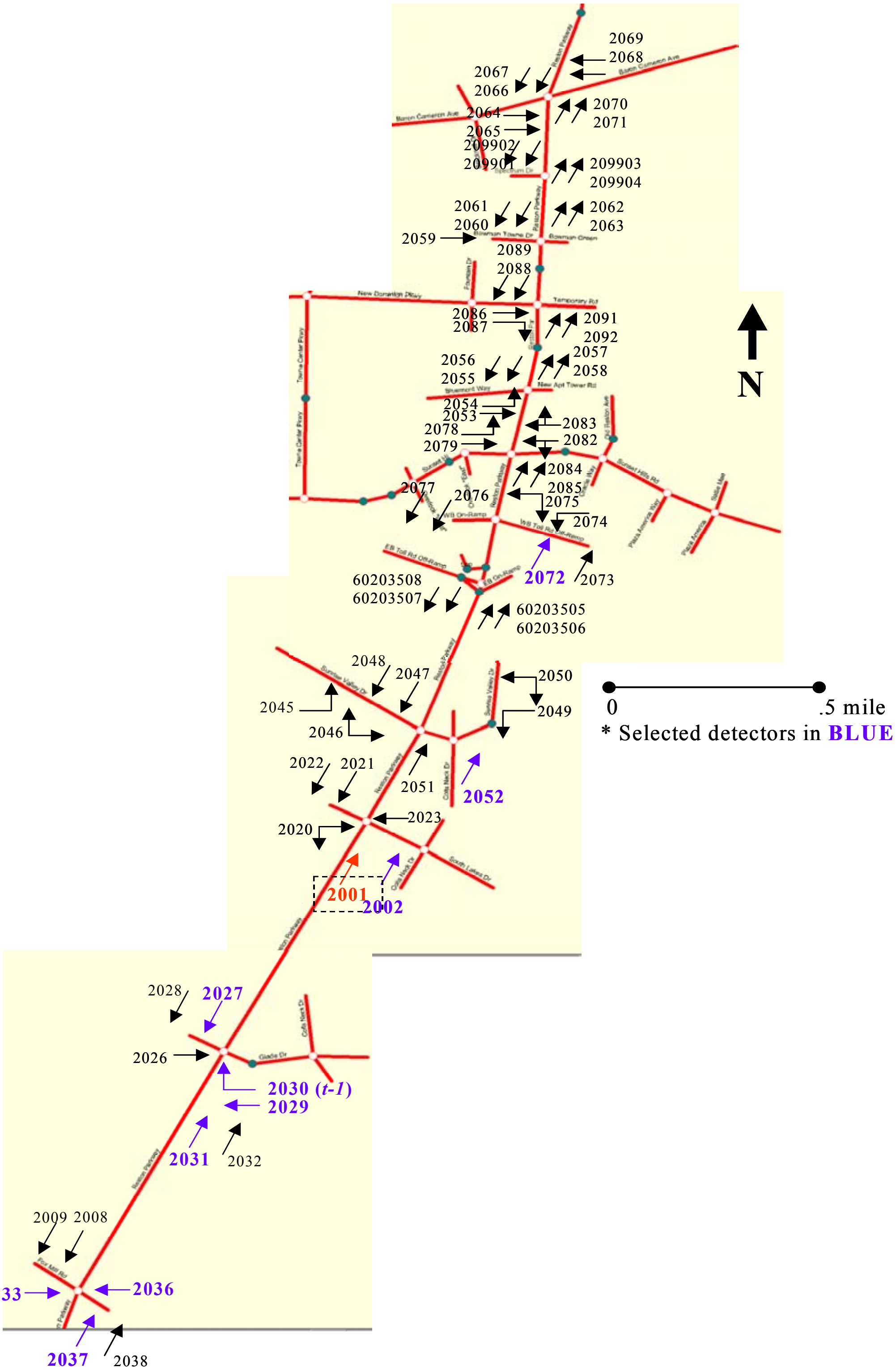


Figure 6-17: C-STARMA Model Selected Inputs for Detector 2001 Volume Using Extended Network

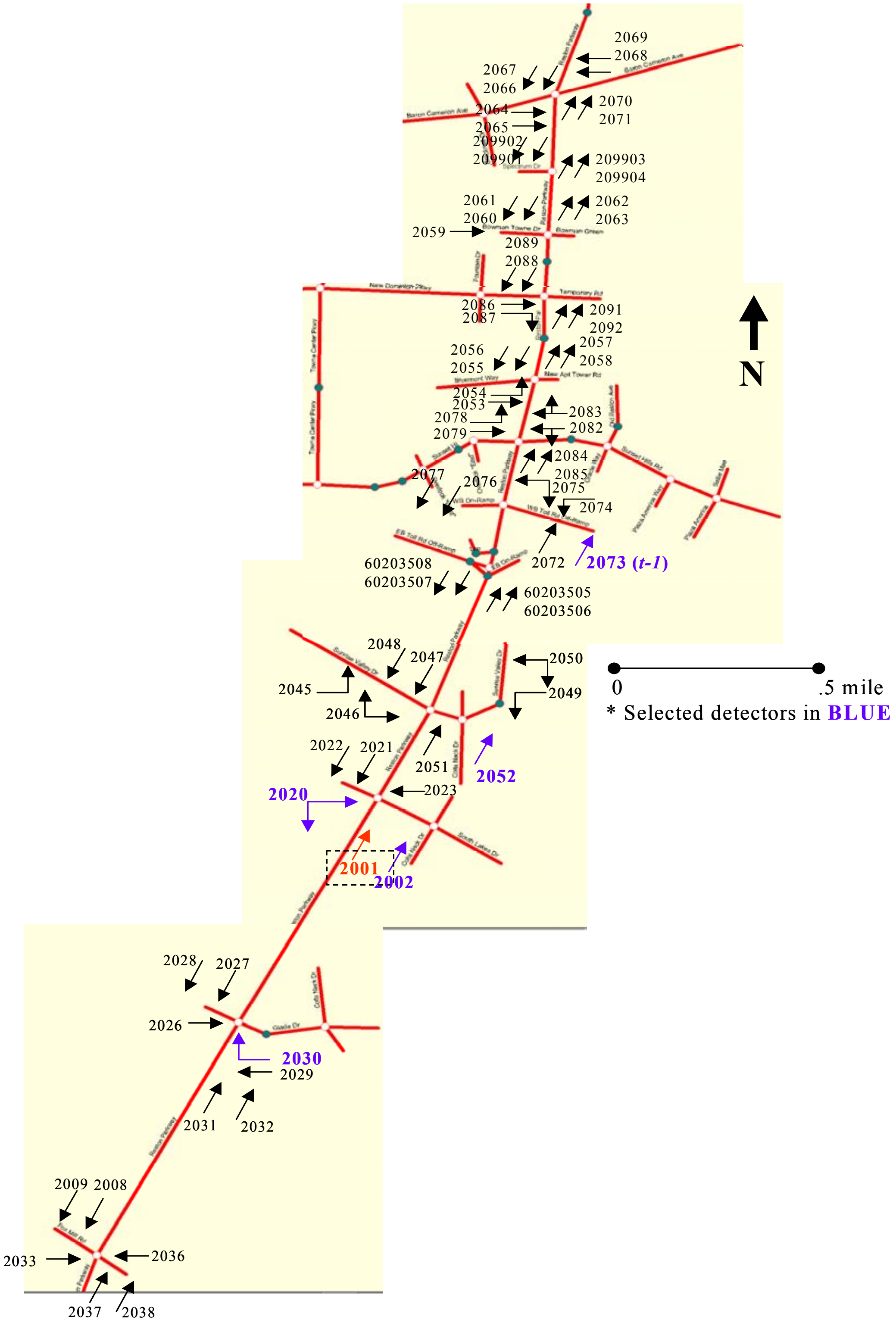


Figure 6-18: C-STARMA Model Selected Inputs for Detector 2001 Occupancy Using Extended Network

6.5 Model Evaluation

We explored six different models to estimate traffic data for the three network scenarios. Table 6-17 highlights the models that produced the best results based upon the mean absolute percentage error metric. The Group and C-STARMA models consistently provided the best estimates of volume, occupancy, and V+KO data based on the MAPE metric. The Group model's final equations were strongly biased towards the C-STARMA model. Other models that performed well included the Regression and STARMA models. The common themes among these models were that they relied upon multi-variate inputs in their model building procedure and that they took into account the underlying spatial relationships among detectors in the arterial network. These particulars models based their imputation on other network detectors as surrogate measures for data at the location of interest. Data from network detectors were utilized to impute data at the location of interest. These models also exploited the spatial correlation among the network's detectors.

Table 6-17: Model producing the best MAPE results

	MODEL SCENARIO			
Metric	**Upstream Detectors**	**Up / Down stream Detectors**	**Downstream Detectors**	**Extended Network Data**
Volume	C-STARMA	STARMA	C-STARMA	Group
Occupancy	Regression	Group	Group	Group
V+KO	Group	STARMA	C-STARMA / Group	Group

Table 6-18 summarizes the model precision at imputing volume and occupancy. A model's precision was judged by the percentage of estimates (residuals) that fell within the +/- 5% range. The results listed validate the utility of the Group and C-STARMA models. The Group model was biased towards the C-STARMA model, and performed well at estimating the traffic data in each case.

Table 6-18: Models producing the largest number of estimations within +/- 5% range

MODEL SCENARIO				
Metric	**Upstream Detectors**	**Up / Down stream Detectors**	**Downstream Detectors**	**Extended Network Data**
Volume	C-STARMA	STARMA	Group	Group
Occupancy	C-STARMA	Group	Group	Group
V+KO	C-STARMA	Group	C-STARMA / Group	C-STARMA

6.5.1 Historical Average

These naïve models performed reasonably well given their simplicity and that fact that signal control systems already calculate these values. They perform well under situations when current the traffic match those of normal historical conditions. However, they were not accurate estimators of missing data in situations where traffic conditions deviate from historical patterns.

6.5.2 Regression

The Regression models based estimations upon traffic data from neighboring detectors of the detector of interest. The model development process was intuitive since the correlation among detectors within the sub-network can be visually identified via the traffic data plots. In essence, the Regression models utilized surrogate sensors as estimators and took advantage of spatial correlations among these network detectors. This model type performed well for estimating traffic data across all four network scenarios. In addition, implementation of this model type for ITS applications should be simple provided there is sufficient availability and proximity of neighboring detectors to the points of interest. The traffic engineer would need to build regression models for every detector in the network and to specify any other detectors that would serve as parameters to each model.

6.5.3 Time Series

Our implementations of this model type were consistently the worst performing estimators of traffic data in all four network scenarios. These models only used time series data from the detector of interest to estimate data for that same detector. The inherent problem with implementing this model for practical purposes is that historical data may not be available. Although our models produced comparable results to those by Williams et al. (1999), we found that our spatially oriented models outperformed univariate time series applications.

6.5.4 STARMA

The Reston Area Network of system detectors was a good candidate application for this model type. The STARMA model extended the time series model by including time series data from surrogate sensors, in addition to the data from the sensor of interest. The STARMA models implemented in this research performed very well at estimating traffic data across the network scenarios. It provided better results than the time series models that used data only from the single detector of interest. The STARMA model's strength lies in that it exploited the spatial correlation among the network's detectors.

6.5.5 C-STARMA

The C-STARMA model was a new development proposed by this research. This model was based upon the STARMA model, but extended that procedure to factor in contemporaneous variables to its model composition. The C-STARMA combined the strengths of the Regression and STARMA models by exploiting contemporaneous data, time series data, and the spatial relationships among the detectors.

This new model type consistently performed well, and outperformed the other models in the majority of the analytical scenarios. We observed that it consistently produced the lowest MAPE scores for imputing traffic data, as well as showing remarkable precision.

The C-STARMA model building procedure was a primary factor to its strong performance at data imputation even in situations when a large set of input variables (detectors) was present. Recall that the model featured a preliminary feature selection step that performed an initial screening of the available input data. This step was critical in eliminating the insignificant data sources and narrowing down the list to only those sources that indicated (spatial) correlation to the location of interest. This benefit was observed when compared to the STARMA model, which accounted for all the initial variables in its model, and then performed backward selection to prune variables. The C-STARMA implemented Stepwise Regression in building its model starting with fewer variables. Ultimately, the C-STARMA models were simpler and contained fewer parameters than those of the STARMA class.

6.5.6 Group

The Group model was implemented using a regression model with the other models as input parameters. As we described, the multi-variate models provided very good estimates of missing traffic data, and the group model expanded upon that premise by providing most of the low MAPE values. The specific equations for each of the network scenarios are listed below. The C-STARMA was consistently given the highest weighting in all of the group models. This was expected since that particularly model used the maximum data from the network at any given interval. The Group model may be a good estimator for research purposes; however, it would be impractical to implement for signal control systems due to the amount of processing involved to generate the input models.

6.5.6.1 Network Scenario 1: Upstream detectors only as model input

Volume regression equation:

2001 Vol = -7.9 + .33 Historical Average + .25 Regression + .01 Time Series
-.01 STARMA + **.45 C-STARMA**

Occupancy regression equation:

2001 Occ = .04 + .08 Historical Average + .05 Regression - .26 Time Series
+ .08 STARMA + **.83 C-STARMA**

6.5.6.2 Network Scenario 2: Both upstream and downstream detectors as model input

Volume regression equation:

2027 Vol = 5.6 + .11 Historical Average + **.83 Regression**

- .01 Time Series - 0.04 STARMA + .12 C-STARMA

Occupancy regression equation:

2027 Occ = .18 + .11 Historical Average + .17 Regression

+ .05 Time Series - .14 STARMA + **.74 C-STARMA**

6.5.6.3 Network Scenario 3: Downstream detectors only as model input

Volume regression equation:

2037 Vol = 1.5 + .18 Historical Average + .08 Regression

- .06 Time Series + .18 STARMA + **.78 C-STARMA**

Occupancy regression equation:

2037 Occ = - .25 + .32 Historical Average + .10 Regression

+ .01 Time Series - .06 STARMA + **.71 C-STARMA**

6.5.6.4 Extended Network Data Model

Volume regression equation:

2001 Vol = -9.1 + .16 Historical Average + .09 Regression - .03 Time Series

+ .06 STARMA + **.72 C-STARMA**

Occupancy regression equation:

2001 Occ = .11 + .07 Historical Average + **.55 Regression** + .004 Time Series

+ .04 STARMA + .34 C-STARMA

6.6 Summary

Due to the underlying spatial relationship among system detectors in the arterial network, the multivariate models that factored in this characteristic performed better than the univariate models at imputing traffic data. The spatially oriented models, such as the Regression, STARMA, C-STARMA, and Group models consistently outperformed the univariate models (Historical Average, Time Series) when evaluated by the mean absolute percentage error metric for volume, occupancy, and V+KO estimations. Furthermore, these models were quite precise by producing a large percentage of estimates that fell within the +/- 15% error range.

Of the spatial models, the C-STARMA was validated to the best performing model in each of the network scenarios. This model type drew from the strengths of the Regression and STARMA models by exploiting the spatial correlation among detectors as well as contemporaneous data. Finally, the C-STARMA model building procedure lead to parsimonious model compositions even in cases of large numbers of input sources.

7.0 CONCLUSION

The objective of this thesis was to develop and evaluate alternative methods to impute missing data that are prevalent in intelligent transportation systems applications. The scope was to estimate missing data from off-line system detectors that capture traffic data (volume and occupancy) on arterial roadways. This research investigated the potential of using surrogate data sources, such as neighboring detectors, to accurately estimate the missing data at an off-line detector. Our assumption was that neighboring detectors possess a spatial correlation that can be applied to estimate traffic data for any given off-line detector.

We implemented both univariate and multivariate models and compared their performance using historical traffic data from the Reston Area Network of system detectors. We observed that models that used maximum data on current traffic conditions performed the best. The models that used data from neighboring detectors as input parameters provided accurate estimations of traffic data for any given location. These models also performed reliably across a range of network scenarios.

7.1 Research Findings

This research was able to demonstrate that available network data sources, such as neighboring detectors, were able to serve as accurate surrogate measures for non-responsive point sources. We developed multivariate models that used neighboring detectors as input variables to estimate traffic data at a given detector location. These models generally performed better than univariate models that relied solely upon pre-existing data from the single detector location.

7.2 Research Contributions

7.2.1 Academic Contributions

The academic contributions from this research were the development and proposal of a new model class for data imputation. The requirements for such a model included the ability to use time series data, as well as contemporaneous data, from multiple data sources. In addition, this model type must be able to exploit the underlying spatial correlation among data sources. The development of the C-STARMA model was a direct response for the need of a robust imputation of traffic data for any specific location when a multitude of surrogate measures were available.

The C-STARMA model was based upon the theory of the classical STARMA model. However, the C-STARMA extended the base class to factor contemporaneous data into the model composition. The STARMA class typically utilized time series data from spatially correlated sites. The C-STARMA model included additional parameters to account for the availability of contemporaneous data.

This research proposed the model building procedure for the C-STARMA model, which included innovations to the base class models. These innovations included an initial step to perform variable reduction (feature selection) and an alternative approach to parameter estimation.

The variable reduction step is the first step performed in the model building procedure prior to determining the C-STARMA model class. This is a critical step to reduce the number of potential model inputs to only those that are significantly correlated to the dependent variable.

Once the model class has been determined, the next step was to estimate the model's parameters. This research proposed the application of Stepwise Regression as the parameter estimation technique for the linear C-STARMA models, which consists of parameters for spatial Regression (contemporaneous data) and autoregressive coefficients. Recall that the base technique (STARMA) applied backward selection of variables, which enters all of the variables in the block in a single step and then removes them one at a time based on removal criteria. Conversely, the C-STARMA applies the Stepwise, or Forward variable selection, which enters the variables in the block one at a time based on entry criteria. Stepwise variable entry and removal examines the variables in the block at each step for entry or removal. Empirically, we observed that this parameter estimation technique produced simpler, more parsimonious models.

This development of the C-STARMA model provides researchers with a powerful technique to accurately account for missing data in current transportation studies, such as travel time estimation, optimal selection and placement of detectors, and traffic prediction algorithms.

7.2.2 Application To ITS Traffic Signal Control

This research investigated the suitability of estimation techniques to account for missing data prevalent in intelligent transportation systems applications. Specifically, univariate and multivariate models were derived to estimate replacement values for missing data from off-line system detectors or traffic databases in support of traffic signal control systems.

The results of this research have proved to be promising to support diverse applications in the intelligent transportation systems domain. Next generation signal control systems that rely upon real-time surveillance would benefit by implementing these estimation techniques as a fault tolerant mechanism. Civil transportation authorities can apply these techniques to support traffic signal systems planning, budgeting, and operations. Furthermore, ITS researchers have additional techniques to account for missing data in transportation databases.

7.2.2.1 Next Generation Signal Control Systems

Next generation signal control systems critically depend upon data input from traffic surveillance devices. Devices range from single-wire loop detectors to video cameras. However, these control systems do not have fault tolerant mechanisms in cases when detectors go off-line.

The techniques investigated by this research support the implementation of fault-tolerant mechanisms to solve this problem. Signal control systems can divert operations that rely upon live surveillance to estimated traffic data based upon on-line network sensors. The available sensors act as surrogate measures and provide a basis for accurate estimation of missing traffic data.

Of the multi-variate models, the regression model would be the quickest and easiest model to implement in signal control systems. This would entail developing a model for each sensor within the network, which includes a feature selection to determine which network sensors are correlated. These models can easily be coded into the signal control system to support traffic responsive or traffic adaptive control mechanisms.

7.2.2.2 Optimal Selection and Placement of Traffic Sensors

These estimation techniques would also benefit civil transportation departments that are considering adding new detection locations or implementing expensive surveillance devices, such as video cameras. Historically, there has been no scientific manner to select the installation location and number of surveillance devices to employ. These authorities can apply the traffic estimation techniques from this research to determine the quality of data from current surveillance devices, and ultimately to determine the optimal location and number of new detectors to deploy.

Traffic engineers can apply this research towards determination of optimal placement of sensors to capture maximum data with few resources (critical sensors). This could feasibly reduce the number of sensors to field and to maintain if system can accurately estimate traffic conditions based upon a few critically situated sensors. Thus, both engineering and financial benefits can be realized from this research.

7.2.2.3 Support of On-Going Traffic Studies

There is a significant amount of on-going research in ITS field that relies upon historical traffic data archived databases that plagued with erroneous or missing data. These techniques provide researchers another means to accurate and reliably substitute for missing data.

Civil transportation authorities, such as the VDOT STSS, rely upon historical as well as current sensor data to evaluate current timing plans and develop new ones. In cases where critical detectors are unavailable, the techniques in this research enable traffic engineers to still get full and accurate depiction of their jurisdiction's traffic condition. In addition, traffic engineers can now more accurately evaluate and compare the application of fixed timing plans against traffic responsive or adaptive control mechanisms.

7.3 Recommendations for Further Research

This research theoretically demonstrated the applicability of imputation techniques to support fault-tolerant mechanisms for real-time signal control systems. Further research can be performed to justify this claim by automating these imputation techniques to estimate missing data in real-time systems. Such an application can be executed in real-time and automates the selection of available detectors as model inputs to estimate traffic data for any given point detector. Further research should simulate traffic responsive or adaptive networks to investigate the performance of these imputation techniques in real-world applications.

Furthermore, our research findings can support a wide range of current ITS research activities that require a full account of traffic data, such as traffic monitoring, traffic control, and data mining.

7.3.1 Assumptions on Automating the C-STARMA Methodology

This section lists a set of assumptions about the C-STARMA modeling procedure to enable its automation for use in real-time ITS (signal control) applications. The C-STARMA model was described as being based upon the Regression and STARMA techniques. It is possible that with enough assumptions to reduce the implementation complexity, the performance of the C-STARMA model may approach that of the Regression and/or STARMA models. However, the benefit is that the implementation is simplified and automation is made feasible without significant analyst intervention during the model diagnosis/development, implementation, and execution phases. These assumptions include, but are not limited to, the following:

- Select system detectors located downstream from the detector to be modeled. For a detector along the main corridor, the selected input sources should be the detectors along the main corridor approximately one to two intersections adjacent to the intersection of interest. It is also appropriate to select upstream detectors only, or a combination of both upstream and downstream detectors as model inputs. In these cases, the spatial order should be limited to two intersections adjacent to the intersection of interest.

- The C-STARMA model can be appropriately limited to linear models, i.e., include only the autoregressive components in addition to the contemporaneous components. For 15-minute data evaluated in this research, the autoregressive time lag component can be set at one, i.e., $z_l(t-1)$ which is the observation of the random variable Z at spatial lag l at time $t-1$.

8.0 REFERENCES

Statistical Methods:

1. Davis, Gary A., Nihan, Nancy L. "Nonparametric Regression and Short-Term Freeway Traffic Forecasting." *Journal of Transportation Engineering*, Vol. 117, No. 2, March/April 1991, pp. 178-188.

2. Greene, William H. *Econometric Analysis*, 3rd Edition. Prentice Hall, New Jersey (1997): pp. 418-445.

3. Kennedy, Ruby L. et al. *Solving Data Mining Problems Through Pattern Recognition*. Prentice Hall, New Jersey (1997): pp. 8.2-8.4.

4. Little, Roderick J. A., Rubin, Donald B. *Statistical Analysis with Missing Data*. John Wiley & Sons, Inc. (1987).

5. Johnson, Dallas E. *Applied Multivariate Methods for Data Analysis*. Duxbury Press (1998): pp. 93-112.

6. Johnson, Richard A., Wichern, Dean W. *Applied Multivariate Statistical Analysis 2nd Edition*. Prentice Hall, New Jersey (1988): pp. 340-371.

7. Mendenhall, William, Sincich, Terry. *Statistics for Engineering and the Sciences*, 4th Edition. Prentice Hall, New Jersey (1995).

8. Mulhern, Francis J., Caprara, Robert J. "A Nearest Neighbor Model for Forecasting Market Response." *International Journal of Forecasting* 10, 1994, pp. 191-207.

9. Pfeifer, Phillip E., Deutsch, Stuart J. “Identification and Interpretation of First Order Space-Time ARMA Models.” *Technometrics*, Vol. 22, No. 3, August 1980, pp. 397-408.

10. Pfeifer, Phillip E., Deutsch, Stuart J. “A Three-Stage Iterative Approach for Space-Time Modeling.” *Technometrics*, Vol. 22, No. 1, February 1980, pp. 35-47.

11. Ripley, B.D. *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge, U.K. (1996): pp. 23, 335, 336.

12. Sadek, Adel W., Demetsky, Michael J., Smith, Brian L. “Case-Based Reasoning for Real-Time Traffic Flow Management.” *Computer-Aided Civil and Infrastructure Engineering* 14, 1999, pp. 347-356.

13. Shumway, Robert H. *Applied Statistical Time Series Analysis*. Prentice Hall, New Jersey (1988).

14. Williams, Billy W., Durvasula, Priya K., Brown, Donald E. “Urban Freeway Traffic Flow Prediction: Application of Seasonal Autoregressive Integrated Moving Average and Exponential Smoothing Models.” *Transportation Research Record 1644*, TRB, National Research Council, Washington, D.C., pp. 132-141.

15. Williams, Billy W. *Modeling and Forecasting Vehicular Traffic Flow as a Seasonal Stochastic Time Series Process*. Ph. D. Dissertation, University of Virginia Department of Civil Engineering, Report No. UVA/29242/CE99/103, June 1999.

16. Yaffee, Robert. *Introduction to Time Series Analysis and Forecasting*. Academic Press, New York (2000).

Data Requirements in Transportation Systems:

17. Dailey, D.J. “A Statistical Algorithm for Estimating Speed from Single Loop Volume and Occupancy Measurements.” *Transportation Research Part B: Methodological*, Vol. 35, No. 5, June 1999, pp. 313-322.

18. Gartner, Nathan H., Stamatiadis, Chronis, and Tarnoff, Philip J. “Development of Advanced Traffic Signal Control Strategies for Intelligent Transportation Systems: Multilevel Design.” *Transportation Research Record 1494*, TRB, National Research Council, Washington, D.C., 1995, pp. 98-105.

19. Gartner, Nathan H., Tarnoff, Philip J., and Andrews, Christina M. “Evaluation of Optimized Policies for Adaptive Control Strategy.” *Transportation Research Record 1324*, TRB, National Research Council, Washington, D.C., 1991, pp. 105-114.

20. McNally, Michael G., Mattingly, Stephen P., Moore, James E., Hu, Hsi-Hwa, MacCarley, C. Arthur, and Jayakrishnan, R. “Evaluation of Anaheim Adaptive Control Field Operational Test.” *Transportation Research Record 1683*, TRB, National Research Council, Washington, D.C., 1999, pp. 67-77.

21. *Traffic Control Systems Handbook*. U.S. Department of Transportation, February 1996.

22. Nam, Do H. and Drew, Donald R. Automatic Measurement of Traffic Variables for Intelligent Transportation Systems Applications, Transportation Research Part B, 1999, pp. 437-457.

23. Skabardonis, Alexander, Bertini, Robert L., and Gallagher, Brian R. Development and Application of Control Strategies for Signalized Intersections in Coordinated Systems, Transportation Research Record No. 1634, 199x, pp. 110-117.

24. Turner, Shawn M., Lomax, Timothy J. *Developing a Travel Time Congestion Index, Transportation Research Board No. 1564*, 1996, pp. 1-10.

Missing Data in Transportation Domain:

25. Gold, David L.; Turner, Shawn M.; Gajewski, Byron J.; Spiegelman, Clifford. *Imputing Missing Values In Its Data Archives For Intervals Under 5 Minutes.* National Research Council (U.S.). Transportation Research Board Meeting (80th: 2001 : Washington, D.C.).

Cost / Benefit Analysis of Surveillance (Detectors):

26. *How to: Operating and Maintaining Traffic Control Systems*, Institute of Transportation Engineers (October 17, 1994).

27. Kraft, Walter H. *NCHRP Synthesis of Highway Practice 245: Traffic Signal Control Systems Maintenance Management Practices*, TRB, National Research Council, Washington, D.C. (1997).

28. Liao, Tsai-Yun, et al. Fuel Consumption Estimation and Optimal Traffic Signal Timing, U.S. Department of Transportation, Report No. SWUTC/98/467312-1, Texas A&M University, College Station, Texas, 1998.

29. Parsonson, P., *NCHRP Synthesis of Highway Practice 114: Management of Traffic Signal Maintenance*, TRB, National Research Council, Washington, D.C. (1984).

30. Patel, Raman. "ITS Operations and Maintenance Issues", *Intelligent Transportation: Serving the User Through Deployment, Proceedings of the 1995 Annual Meeting of ITS America*, 1995.

31. Sadek, Adel W., Smith, Brian L., and Demetsky, Michael J. *Artificial Intelligence-Based Architecture for Real-Time Traffic Flow Management*, Transportation Research Record No. 1651, 1998, pp. 53-58.

32. Skabardonis, Alexander, Gallagher, Brian R., and Patel, Kartik P. *Determining Capacity Benefits of Real-Time Signal Control at an Intersection*, Transportation Research Report No. 1683, 1999, pp. 78-83.

33. *Traffic Signal Maintenance: An Urban Consortium Information Bulletin*, Public Technology, Inc. (June 1982).

34. Highway Capacity Manual Special Report 209, Third Edition, Transportation Research Board, National Research Council, Washington, D.C. 1998.

APPENDIX A: C-STARMA MODEL IMPLEMENTATION PROCEDURE

This appendix describes a step-by-step procedure for automating the C-STARMA model for real-time ITS applications to impute 15-minute traffic data (volume and occupancy). The following procedure takes into account the assumptions described in Section 7.3 to enable the automation of the C-STARMA methodology without significant analyst intervention during the model diagnosis/development, implementation, and execution phases. These assumptions lead to an abridged version of the C-STARMA model building methodology suitable for automation.

1. Collect Data:

   Collect 15-minute arterial network (system detector) data for the detector of interest, and from detectors at the same intersection. Collect data from system detectors at adjacent intersections up to two intersections away (along the main corridor).

2. Perform Model Identification:

   The model should specify the following variables as candidate input parameters:

   - Contemporaneous Variables:
     - $z_l(t)$ where *l* is the spatial lag indicator which specifies data at time interval *t* for neighboring detectors,
   - Autoregressive Variables:
     - $z_i(t-1)$ the autoregressive parameter at the location of interest,
     - $z_l(t-1)$ the autoregressive parameters for neighboring detectors.

3. Estimate Model Parameters:

   Perform a Stepwise Regression (Linear Least Squares) upon the specified input variables. This step down-selects the significant model input variables and estimates their parameter coefficients for the linear model.

4. Perform Model Diagnosis:

   Execute the derived model to impute traffic data at the point of interest. Compare the estimated values against the actual values. The Mean Absolute Percentage Error (MAPE) is a suitable statistic to evaluate model performance. If the model is determined to perform sufficiently well, then it can be applied to impute traffic data for real-time ITS applications. However, if the model is insufficient, then the analyst will need to revert back to Step 2 above and select detectors (inputs) from alternative network scenarios (e.g., downstream-only, upstream-only, or both up-/down-stream detectors).