Xiao Xiao 2015

Estimation of missing flow at junctions using control plan and floating car data

Cover figure comes from http://www.northeastern.edu/datascience/

SUPERVISING COMMITTEE:

Prof. Dr. Ir. S.P. Hoogendoorn

Dr. Yusen Chen (TNO)

Dr. Yufei Yuan

A.Jamshidnejad, MSc

Prof. Dr. Em. Bart De Schutter

Preface

The thesis is for the degree of Master of Science (MSc) in Civil Engineering, with a specialization in Transport, Infrastructure and Logistics (TIL) from University of Technology Delft. The research for the thesis is under the supervision of the graduation committee, cooperating with TNO (Netherlands Organisation for Applied Scientific).

First and foremost, I wish to thank each of the members of my committee for their advices and the time they spent helping me. These include: Prof. Serge Hoogendoorn, for his always very useful criticisms and the way in which benefitting from his erudition has helped me see the bigger picture; Ana Jamshidnejad, who never failed to show up at meetings and offer her comments on my report; and thank Prof. Em. Bart De Schutter for his support and being in my committee. My daily research supervisors, Dr. Yusen Chen and Dr. Yufei, they are so helpful and so generous with their time. I always kept in mind their dicta: “Be focused and concentrated” Dr. Chen would say; “Be precise and patient, and look deeply into the results”, Dr. Yuan urged me. I realized that, in science, exciting innovations come from diligent and patient labor.

Secondly, I wish to thank all my colleagues at TNO; they are all friendly and kind with me. Dr. Taoufik Bekri gave me suggestions and help about dealing with big data.

Finally, I thank all my friends and relatives, who inspired me with confidence, and gave me an opportunity to feel needed. I especially would like to thank my parents, without whose support I could not have enjoyed the opportunity to do this work, or the pleasure of engaging it while meeting so many interesting scholars and scientists, and workers in related fields, around the world.

These teachers, colleagues, and friends, and the work their community has enabled me, I hope, in this thesis make some contribution that I hope will be harbingers of others to come. To recognize that science and scholarship are not only, nor primarily, the acquisition of useful information; they are forms of that most mysteriously difficult of human endeavours, and also, to be sure, among the most satisfying: the labour of thinking.

Summary

In this thesis, the author develops a series of methods for estimating missing traffic flows at urban junctions (intersections) through direct observations, and the combined use of multiple data sources with data fusion.

In the fields of Dynamic Traffic Management (DTM) and Intelligent Transportation Systems (ITS), data acquisition has been central and important. Since, in a field such as traffic systems (along with many others), a reliable and complete database is a basis for management and control of operations; if the data are unreliable or even partly missing, there will be trouble in the traffic systems. In the SCATS (Sydney Coordinated Adaptive Traffic System), traffic flow data are often found missing from loop detectors. The data quality of traffic flows directly affects the efficiency of an adaptive control system. Therefore, estimation of these missing flow is important. The research question pertinent to this addressed in this thesis is: What are the best ways to estimate the missing flows at an urban junction?

The author puts this question by doing research using two sources: data from SCATS and FCD (floating car data) obtained from taxis. SCATS makes available, in addition to the raw data of traffic flow observations, control plan (timing) information. One way to estimate the missing data is by using available traffic flow observations. Another possibility is to use data from "missing" sources, along with traffic flow theory, for data fusion is also a possible method. Four main approaches are possible based on these two methods. The author tests each one by leaving aside actual detected values in order to represent a missing flow. The evaluations are made based on two kinds of target: values that are detected in each detection period, and ground-truth values. The original data represent the former, while more refined data are used to represent the latter.

The first approach involves considering the historical pattern of data. The algorithms in this approach are based on the retrieval of existing observations, and the use them to estimate the missing flows. There are two ways of developing these algorithms. The first is by using a fixed detector. In this case, the flow observations are seen as independent variables, and possible flows that have not been detected are related to the flows observed in the recent past (through online or offline analysis). The second way is to consider the flow and the green at the same time, taking the flow/green ratio as an independent variable, and making the same assumptions as them to the flows. When it comes to application, the flows presented in the historical data are representative. However, the flow/green ratio fails to take effect. The reason is that the green is not adaptive to flows – the green does not change due to the flows for most of the time in the given data. In the results, even for days with stable flows, the estimated values using this approach may "shift" a little bit during a specific period of a day from the actual values. Compared with other approaches, the general

performance is the best when dealing with smoothed data.

The second approach considers the spatial distribution of flows over lanes. The approach uses algorithms to refer to observations from detectors in other relevant lanes (for example, in the same phase) during the same period. This is done by comparing each lane with each other lane that is relevant to obtaining reference values, and then weighting those values according to their spatial characteristics. In application, the flows from lanes in the same phase are representative. In the results, this approach captures the instant changes of flow during short periods. Besides, it involves low relative errors when dealing with both original raw data and processed (smoothed) data.

The third approach tries to link FCD with flows using the data fusion concept. There are two ways to consider this approach. The first is to use the relation of FCD speeds and loop flows. In traffic flow theory, a fundamental diagram of speed and flow is applicable under certain conditions on a freeway. The assumption is also made that there is a certain relation between the speed and flow of the traffic stream at junctions. Although the traffic speed is unknown, the speed from FCD is used to represent that of the whole stream. In the application, the relations show a general negative influence of flow on average FCD speed. However, the linear extrapolate tools do not so far yield uniformed fitting curves. The specific parameters vary from stream to stream. The estimated flows involve larger errors than in other approaches. The second part of this approach uses the relation of FCD counts and loop flows. The counts of FCD show the numbers of taxi streams, which represent part of the total traffic stream. There is a roughly positive relationship between FCD counts and total traffic volumes. A linear curve is made to fit for this relation. In the application, this provides an estimation of traffic on a stream level, with large errors. However, the errors are slightly smaller than when using the speed-flow relation.

The fourth approach performs the estimation by applying a multiple linear regression. In the first step, the potential contributions of observations to each other are assumed, and the parameters showing these contributions are calibrated. In the second step, all these parameters are applied to the available values to calculate the missing values. The input range and analysis interval are both key factors influencing its performance. A suitable trade-off should be made such that the inputs are relevant enough, and the analysis interval not too short. In the application, MLR (A4) has the best record of dealing original flow data. Its performance depends on the amount of actual flow.
In addition to these four approaches, two methods are developed for integrations. They are expected to further improve the estimation.

The first integration uses an iteration method. The iteration goes between the values calculated using historical flow patterns and those based on the spatial flow distribution. This done by filling in the missing values using both approaches, and updating the values by applying the updating weighting factors and cross-comparisons.

The results indicate that when one of the other approaches does not perform well, the iterative one tends to be the better-performed alternative. When the performances of two approaches are close, the iteration can provide a better estimation than either one of them. Since there is no way of knowing when one approach will perform better than another one, using an iterative approach is a reliable and safe solution. It also requires fewer direct observations than the MLR integration method. The only drawback in this case is the computation costs, which are relatively high.

The second integration starts from the MLR approach, and considers the information from the historical pattern and lane flow spatial distribution in order to improve the relevance of inputs. Those results turn out to be best in which the actual flow is relatively large, while and worse than the integration using iteration when the actual flows are low. Fortunately, it produces results faster than the integration using iteration.

The data processing methods in this thesis include the coding of junctions, the extraction and retrieval of flows and timing plan, and the processing of FCD trajectories. Some tools are developed for achieving these processes, include a new coordinate system for forming FCD trajectories at urban junctions and an application of the PLSB (Piecewise Linear Speed Based) trajectory method.

In discussing the experiments and case studies, the author sets up the experiment by defining some of the key influential factors observed in the initial experiments. Two junctions were chosen to apply the calibration and validation, respectively. Different approaches and methods were applied to the first junction, under different scenarios for the tests and calibration. In the first two approaches, the primary results using the initial methods, and improved results using updated formulas are included. Then, two integrated methods are applied to the second junction for the purpose of validation, which gives confidence to the extension.

In the end, a comprehensive comparison is made among approaches and methods. The grand conclusion is that the best methods for estimating missing flows at the urban junction are determined according to the specific situation. Except for error indicators, criteria such as the data needed, the number of direct observations, and the computation costs should also be considered. Even for error indicators, the methods have different advantages with regard to specific missing data type (short- or long-term) or data input type (raw data or processed data). The methods used here are recommended for practical application to estimating missing flows at urban junctions under stable traffic conditions, in situations meeting the criteria identified in this thesis.

Table of contents

List of Figures

List of Tables

1. Introduction

This chapter provides background to the research presented in this thesis, and identifies the main research question and the thesis's contributions. Section 1.1 provides background on the widespread phenomena of missing data in traffic systems. Major causes of this missing data are introduced, followed by their consequences. Relevant research conducted in similar areas is presented and classified. Based on current problems, the main research question is raised in section 1.2, followed by several sub-questions. In section 1.3, a complete induction-deduction loop is introduced to explain the research process used herein. Section 1.4 explains the contributions of the thesis, both practically and scientifically. Finally, section 1.5 provides an overview of the thesis structure.

1.1. Background

Traffic congestion is a major urban problem. To address it, solutions such as dynamic traffic management (DTM), or traffic systems such as Intelligent Transportation Systems (ITS) have been developed. Traffic flow data play a significant role in accurate traffic state estimation and efficient traffic management. Hence, the availability and quality of traffic flow data are of great importance, and making estimations of traffic flows is needed for both on-line (real-time) and off-line processes. Meanwhile, these processes all rely heavily on the data. For example, among traffic systems, some use adaptive signal controls in response to the changing traffic volumes in urban regions, and these adaptive systems develop their control plans based on the real-time flows they have gained. SCATS is a worldwide-used intelligent transportation system, which determines its control plan based on the traffic volume in a given unit of time. If the data are incomplete or even completely missing, the operation of the system will be difficult. Then, it becomes important to improve the traffic flow data quality by estimating the missing flow.

As described above, the position of data acquisition is crucial. Traffic data are collected in large quantities by various sensors and in multiple ways in intelligent transportation systems (ITS), including loop detector, GPS, video, Bluetooth, and others. Loop detector is one of the most important measurements for collecting traffic data. Many means have been tried of improving the loop data quality. The detector event data-collection (DEDAC) system was developed by the TransNow research team at the University of Washington. This system combines digital data-collection techniques, a multimedia high-resolution timer, and multithreaded programming techniques. The system can do real-time loop data quality evaluation, loop malfunction identification, and loop error correction. Zhang et al. (2003) developed a new dual-loop algorithm by conducting various checks to test the validity of individual vehicle data. However, this kind of system may be too expensive for most government agencies, especially in developing countries. There are numerous issues related to installation, comparability, and maintenance. Therefore, most of the existing traffic systems, without specific tools to guarantee their data quality, still need efforts to solve their problems of missing data.

Missing data: causes and current situations

Clearly, data missing are problematic for any functions that detect data that are to be used in a dynamic traffic management or other ITS system.

There are multiple reasons for data missing this phenomena. The consequences, of course, are serious. According to Boyles (2011), causes of missing data include data detector failures, failures in power or communication, and man-made factors (Tang et al. 2015) such as incorrect observations are also included. Data transfer is another

trigger, as there can be loss of data packages during transmission (Qu et al. 2009).

Missing data is widespread in many traffic systems. For example, Turner et al. (2000) found that, in archived data from San Antonio, Texas, nearly 25% of data records are missing or unreliable. Nguyen and Scherer (2003) note that 25-30% of the detectors from the Virginia Department of Transportation default permanently. In addition to the defaulting of detectors, some functioning detectors also provide unreliable or missing values. Kwon (2004) suggests around of the data for functioning detectors is missing. Qu et al. (2009) say that in Beijing the missing ratio in ITS of daily traffic flow volume data is around 10%. Four percent of this is due to the malfunctioning of detectors, and 6% to other reasons. In the case of specific detectors, some even show a missing ratio between 20% and 25%.

The missing flow at a SCATS junction

A considerable amount of research has been conducted on traffic flows on freeways. Yet, there is a need for more research on traffic flow at urban junctions.

Smith et al. (2001) suggest that traffic signal systems represent the first widespread deployment of ITS. SCATS is one of the adaptive traffic control systems at junctions that adapt control schemes according to traffic flow volume. Stevanovic (2012) shows that the long-term benefits of SCATS include better performances than some other plans. However, due to the device failure, the data quality in the system is low.

To improve the data quality at junctions in SCATS, a better understanding of flows is needed. For example, the flows in lanes vary according to streams, and turning or timing groups. While SCATS does not compare the flow counts over lanes, this study helps to look at this.

Relevant studies using SCATS data

Some studies have been conducted using similar data sources, and the data fusion of FCD and loop data from SCATS has been used by researchers to estimate traffic states or make predictions. For example, Chen et al. (2013) built a dynamic simulation model using SCATS and FCD from Changsha for the assessment and evaluation of urban networks. Li et al. (2014) use FCD to apply fusion methods for travel time monitoring in urban areas. Zheng et al. (2012) complete link travel times by using sparse probe vehicle data. Lu et al. (2012) compare the counts obtained from video recordings with flows in SCATS.

1.2. Research question

Previous studies have sought to improve the quality and comprehensiveness of raw observation data from monitoring systems. The focus in this thesis is on urban signal-

controlled intersections. The main research question then is:

What are the best ways to estimate traffic flow volumes missing at a junction?

This research question directs our thinking on: whether the traffic flows are collected well at detectors or not, and whether the values detected are consistent with each other. Therefore, to specify the research question, several sub-questions are formulated, as follows:

- *What is the data quality at junctions in a traffic control system such as SCATS?*
- *How to check the flow consistency of a junction?*
- *How to connect FCD with loop data to provide more information ?*
- *What are the differences in methods in estimating missing traffic flows?*

1.3. Research process

The research makes use of an induction-deduction loop concept. First, from the induction side, the author identifies the phenomena of missing flow by analyzing the flow data from SCATS (Chapter 3). A research question is raised (Chapter 1). To find out the answers to the research question, the author undertakes actions from both the induction and deduction side. Then the findings from one side act as the inputs to another side, iteratively.

First, from the induction branch on the right hand side, the observed flows are analyzed from available data sources, and their patterns are observed. The author makes some tentative hypothesis about the relationships between the missing flow and other observations (Chapter 2). Primary approaches are made based on these relations: Using the nearest detector or the average of neighboring days (as close to time or distance as possible) (Chapter 4). These primary approaches yield some initial results (Chapter 6). These processes are then carried out in the induction branch, as shown in Figure 1 1.

A literature review is conducted based in part on the deduction branch on the left hand side (Chapter 2). The primary approaches used in the induction section are combined with theories and concepts from previous studies (Chapter 4). This leads to the suggestion of new approaches, whose performances are tested in case studies (Chapter 6).The updated approaches are compared with various observations, and the research process goes back to induction branch.

After several loops of switching between induction and deduction processes, feeding the inputs of induction or deduction with outputs from both, the methods with better results will be confirmed and applied for a validation. The methods still with unsatisfactory performance even after many trials are reserved for use as a reference. The section on the final output (Chapter 7) presents some suitable means of missing

flow estimation, provides recommendations for application in practice, and points out possibilities for improvements to ongoing research.

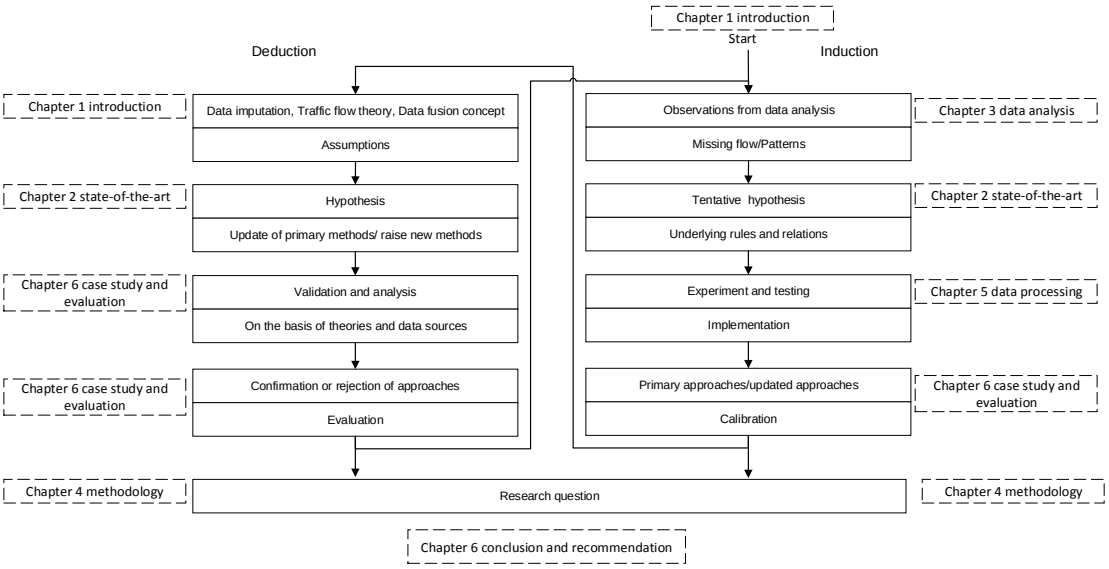


Figure 1-1 flow chart for the research for the thesis

In conclusion, it should be noted that the actual research process is far more complex than is suggested by the contents of this thesis. The gains include cumulative outputs from the loops between induction and deduction processes.

1.4. Contribution

Theoretical contribution

The theoretical contributions of this thesis include the development of algorithms, and detailed analysis.

The first contribution concerns the quality of the traffic data. This thesis provides a further look into the flows at an urban junction, and develops a series of methods to improve the data quality at junctions. These methods are easy to apply and suitable for further improvements under various conditions. The research also benefits from traffic state estimation, prediction, management, and control.

The second contribution concerns data. The thesis solves missing flow estimation problems by linking both imputation and data fusion concepts, in considering multiple factors that have been involved in this problem. The trials that link different data sources such as loop flow and FCD have provided some important findings and implications for further research on this topic.

In addition to algorithm development and data, the research for this thesis has provided comprehensive comparisons between different techniques of flow estimation– from simple to complex, from a single aspect to multiple aspects. It also takes into

account factors influencing the performance of algorithms, such as suitable periods, and analysis interval. These detailed factors have usually been ignored or omitted in previous research.

Practical contribution

The practical contributions of this thesis are mainly in flow estimation and data processing:

From the SCATS side, the thesis provides an inspection into the data availability of the whole SCATS network in Changsha, China, showing the data availability of each lane of each junction. This can be a significant information that SCATS can use for purposes of management and control.

In data processing, by sorting and classifying, flow data from SCATS are well organized into groups. Each flow record in the network can be extracted by referring the junction number, date, and lane number. The thesis proposes some solutions in data processing of FCD, too. FCD is classified according to their attributes such as ID, and for each vehicle their actual trip at specific time periods are defined. The author uses a way to form FCD trajectories and to calculate the speed and counts.

1.5. Overview

This section provides an overview of the thesis structure. In this chapter, the background of the research is explained (1.1), followed by a presentation of the research questions (1.2). The main research question involves the concepts of induction and deduction loops (1.3). Contributions are also stated (1.4), followed by an overview (1.5).

Chapter 2 presents a literature review of previous studies on missing flow type (2.1), and missing data imputation methods (2.2). Then, literature on data fusion and other topics are presented (2.3). The author then sketch the argument of the thesis (2.4), followed by a conclusion (2.5).

Chapter 3 mainly looks into the data itself. In 3.1, the quality of data flows from the loop detector in SCATS is analyzed. In 3.2, patterns from the time and space dimensions are shown. The timing plan (3.3) and FCD (3.4) are then analyzed. For FCD, the analysis starts with the data structure and then considers the utilization of speed and counts.

Chapter 4 presents the methodology. First, the author describes the way chosen to express the flow in the general framework (4.1). Secondly, four individual approaches are introduced in 4.2. The first approach (4.2.1) contains two sub-approaches: the historical flow pattern (4.2.1.1) and the historical timing-flow pattern (4.2.1.2). The

second approach is that of lane spatial distribution (4.2.2). The FCD data flow fusion (4.2.3) contains two sub-approaches: they using the speed/flow relation (4.2.3.1) and counts/flow relation (4.2.3.2), respectively. The multiple linear regression is described in 4.2.4. The combination of approaches is presented in 4.3. Two of them are iteration (4.3.1) and advanced multiple linear regression (4.3.2).

Chapter 5 is a short chapter describing the data processes. Two raw data sources are processed: data from SCATS (5.1), containing a loop flow and timing plan, and taxi data from FCD (5.2).

Chapter 6 shows the experiments and the results, including the experiment set-up (6.1), tests and calibration and validation. A comparison and a final evaluation are presented in the conclusion (6.9). Among individual cases, sub-approaches are also tested for the historical flow pattern and the historical timing-flow pattern; as well as the speed-flow relation and counts-flow relation.

Chapter 7 provides the overall conclusions, recommendations for application, and suggestions for future research work.

2. The state-of-the-art

This chapter introduces the state-of-the-art of missing flow estimation as well as the positioning and the reasoning of the thesis work. The literature reviews and previous studies provide theoretical support for the development of methods. Firstly, the chapter gives the general classification of missing data in section 2.1. Secondly, section 2.2 goes through the background of missing data imputation and classifies the current methods. Thirdly, section 2.3 gives brief descriptions about data fusion and other relative topics. Fourthly, section 2.4 shows possible challenges, according to current studies, and positions the research work. The author also describes the plans to face these challenges, which shows the reasoning of the methods. Finally, a conclusion is made in section 2.5.

2.1. Missing Data type

Rubin (1976) gives the definitions for the classification of the missing data. This classification is 'missing data mechanisms'. There are three types of missing data: Missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR).

Among these types of data missing, MCAR shows data that are missing completely at random. MAR shows missing at random, but the missingness of which is not random, and the missingness can be fully accounted where there is complete information. That is to say, the observed values have the same statistical distribution as the other observations. Missing not at random (MNAR) is neither MAR nor MCAR; in this case, the reasons of missing are related to the whether or not there exist a missingness, so the distribution has been hidden.

Usually, it is impossible to identify from a dataset which category they belong to. However, due to the setting up of the sensors and the mechanisms of the detection, it can be assumed that, the missing data in a traffic area belong to the MAR or MCAR. Actually, imputation methods are widely applied according to the relations of missing data or the missingness.

However, traditional imputation methods may face challenges: If there is not enough available value in the dataset, it is hard to get the relations by the few data that are not missing. In this case, some more information is needed. Otherwise, more advanced algorithms are required. In this case, the information from other sources is concerned. This then falls to the same concept of data fusion. The following two parts will give the state-of-the-art for both imputation and data fusion.

2.2. Missing Data imputation

As talked before, the quality of the data acquisition and the analysis of a traffic management system are affected by the degree of missing traffic flow data. Completing the missing flow data is a fundamental step.

Albright, D. (1991) gives the history of Traffic volume estimation and evaluation in the US: During the 1930s, traffic volumes were extensively manually counted. In the 1940s, the measurements were transferred to mechanical ones, which made approaches for data integrity. From the 1950s to 1960s, annual traffic summary statistics got theoretically developed. Historical assumptions, normal distribution of traffic, traffic variability, and data imputation and smoothing were developed and widely used. Many imputation-based procedures were developed for recovering missing data. Some that have been used or related to missing data estimation are presented in the following Table 2-1.

Single imputation and multiple imputation

There are multiple types and several ways of categories for imputation techniques. One way to make classification is single imputation and multiple imputation. Single imputation is to fill in one single value by the processes. One example of the single imputation is "hot-deck". The term "hot deck" indicates that the information comes from the same dataset as the recipients. Sande (1996) uses hot deck imputation, and the units recorded in the sample are replaced by the values obtained from the nearest data record.

Multiple imputation (MI) is to reconstruct or train multiple missing points at the same time (Schafer, 1997). Rubin (1987) develops the method that averages the outcomes across multiple imputed data sets. Nguyen and Scherer (2003) use multiple imputation techniques to account for missing data in support of intelligent transportation systems applications. In the method, each data set imputed is analyzed separately, the results are made average. The standard error term is considered according to the variances of each data set.

Temporal/spatial algorithms

The temporal imputation stands for the algorithms that making an estimation based on the average value of historical data in the same time interval to interpolate missing data. Guo et al. (2008) state that data collection along the time dimension is a fundamental determinant of the nature and utility of the data streams. Many research solves missing data by considering this aspect (Chen and Shao, 2000) (Nguyen and Scherer, 2003), especially when there are no other neighbor detectors. This method has some advanced versions which with the same concept, but equipped with higher techniques, for example, Qu et al. (2009) develops a probabilistic principal component analysis (PPCA) to impute the missing flow volume data based on historical data mining.

Nearest neighbor imputation is one of the hot deck methods which belongs to single imputation methods mentioned in previous part. It estimates the missing value using the average data from one or more of neighboring detectors. The Lane distribution method is one case of this algorithm (Conklin and Smith 2002).

Regression

Regression is another tool that has been widely used. Rubin (1987) uses the regression method for missing data imputation. Pawlak (1993) uses regression imputation by giving a function and estimate the parameters based on the known variables. Chen et al. (2003) apply a regression imputation to filter data from single-loop systems. This method considers the relations relative locations, which is suitable for both urban and

freeway networks. Nguyen and Scherer (2003) use linear regression model, and Al-Deek and Chandra (2004) suggest the model to use nearby detectors. Yuan. Y et.al (2012) uses multi-linear regression to estimate multi-class and multi-lane flow counts from generic freeway surveillance systems.

Except for imputation, regression approach has been widely used in the research related to the analysis of traffic flows. Romana (1999) uses linear regression to form the direct ratio between travel speeds of passing and passed vehicles. Guan et al. (1999) explore the relationship between capacity reduction of the high-occupancy-vehicle (HOV) lane in ingress and egress section. They analyze the impact factors, by a database developed from field collection using both linear regression and non-linear regression model. Fazio, J. et al. (1999) uses correlation coefficients to find the factors that impact fatalities. Smith, B. (2001) uses cluster analysis to group together similar samples of traffic volume conditions to identify intervals of time of day (TOD) signal timing control. Wang, X. et al. (2009) develops Kriging-based methods for mining network and count data over time and space.

Others

Other techniques have also been used in this area. For example, Southworth et al. (1989) introduce RTMAS, which applies time series model to make a prediction as well as the missing value estimation. Wall et al. (2003) present a time-series algorithm for correcting errors in the freeway traffic management system archived loop data. Other similar applications are Monte Carlo techniques (Gelfand and Smith, 1990) (Gilks et al. 1996) (Schafer 1997). Zhong et al. (2004) apply an advanced model based on a genetic algorithm for the missing count estimation. Tang. J. et al. (2015) use a fuzzy c-means (FCM) to impute missing traffic volume data in loop detector and optimize the parameter of cluster size and the weighting factor in FCM model using a genetic algorithm (GA). Some studies are combined and improved gradually. For example, Tanner and Wong (1987) use the data augmentation DA method. Lavori et al. 1995 give the expectation maximization method. In the EM method, the historical average, especially during the day time, is used. Smith and Babiceana (2004) apply a Two-tiered approach, which is known as EM/DA, it makes the combination of the expectation maximization method (EM) and Data augmentation (DA), by adjusting and adding punishes to the imputes according to the period– during the day or the night.

Table 2-1 the framework for the imputation methods for missing flow data estimation

Method categories	Methods	Reference studies
Single Imputation	Hot deck	Sande (1996)
Multiple imputation(MI) and Linear regression	Linear Regression	Rubin (1987) Pawlak (1993) Chandra and A1-Deek (2004) Yuan et al. (2012)
	The propensity score method	Rosenbaum and Rubin (1983)
	The expectation maximization method (EM)	Dempster et al. (1977) Lavori et al. (1995) Schafer (1997) Smith and Babiceana (2004)
	Data augmentation (DA)	Tanner and Wong (1987) Smith and Babiceana (2004)
	Others (Temporal/spatial algorithms, traffic flow theory)	Rubin (1987) Schafer (1997) Chen and Shao (2000) Treiber and Helbing (2002) Conklin and Smith (2002) Huang and Zhu (2002) Nguyen and Scherer (2003) Chen et al. (2003) Ni, D. et al. (2005) Van Lint and Hoogendoorn (2009)
Others	Time series (e.g. ARIMA)	Southworth et al. (1989) Nihan (1997) Wall et al. (2003)
	Monte Carlo techniques	Gelfand and Smith (1990) Gilks et al. (1996) Schafer (1997)
	Genetic algorithms (GA)	Zhong et al. (2004) Tang. J. et al. (2015)
	Probabilistic principal component analysis (PPCA)	Qu et al. (2009)

Traffic flow theories are also considered to support the estimation of missing data. Forexample, Treiber and Helbing (2002) develop an adaptive smoothing method based on the notions from the first-order traffic flow theory, to reconstruct and clean flow observations from dual-loop systems. This approach has been further generalized by Van Lint and Hoogendoorn (2009) to fuse multiple data sources. Except for the methods mentioned that can be classified, there are also many other methods that have been used to make the estimation of missing data, which are not so easy to classify. For example semi-parametric methods (Lawless and Kalbfleisch, 1999),

Bayesian methods (Fitzgerald, 1999) and pseudo-nearest-neighbor approach Huang and Zhu (2002). Boyles, S. (2011) makes a comparison of all the interpolation methods for missing traffic volume data.

2.3. Data fusion and other sources

The previous part has given the fundamental concepts about missing data imputation methods. Although section 2.2 also contains data fusion concept and data sources, this part shows some specific looks into data fusion and some other topics associated with the flow estimation concerned in this research.

Data fusion

The data fusion concept act as an important role in estimating the missing flow due to its special position and utility. Data fusion has been discussed for long and widely used in any offline or online traffic management or data archival system. Currently, traffic data are collected from various sensors such as loop detectors probe vehicles, video cameras, mobile phones and Bluetooth, etc. However, some provide highly correlated data, their data are of different types, with uneven frequency and density. By data fusion techniques, the information from these sources can be made to compliment for each other. Similarly, if data are missing, the relation or the information can also be gained in this way to make consistent and reliable estimating for missing values; especially when there are not enough imputation possibilities provided too few direct observations from a same system.

Linn and Hall's (1991) give a simple three-level model for data fusion, which is also explained by Varshney (1997). In the model, each level has its particular function and purpose; the higher level data fusion is supported by the results from low level, and each level has its corresponding suitable methods. Five general, goal-oriented, data fusion methods are spread in the three levels: data association, positional estimation, identity fusion, pattern recognition, and artificial intelligence (Linn & Hall, 1991). For the first level, the main task is to process raw data, and the outputs are the foundation or inputs for the usage of the higher level. For the second level, further information such as features and patterns are provided;methods like Regression or Neural networks are applied. For the third level, assessments and decisions are made. A table is shown for common data fusion techniques.

Table 2-2 Common data fusion techniques (Linn & Hall, 1991)

Level	General Methods	Techniques
Level 1	Data association	Figure of merit (FOM)
	Position estimation	Kalman filter
Level 2	Identity fusion	Neural network
	Pattern recognition	Cluster analysis
Level 3	Artificial intelligence	Expert system/ Fuzzy logic

As the development of data fusion techniques, the concept of data fusion has been widely and commonly used in the traffic area. The goals and the implementations are updating. For instance, Ou (2011) describe a 'core' and a 'shell' as the two components in traffic data fusion. The core represents the physical laws or assumptions in traffic theory, and the shell stands for the assimilation tools. He uses data-driven assimilation tools to develop a series of data fusion methods by linking the underlying relations of data and checking the consistency. Similarly, this research also checks consistency of data to support the imputation tools.

GPS application

The use of Global Positioning System (GPS) technologies has performed as an important traffic data collection means for all kinds of transportation studies. Li et al. (2002) investigate the minimum sample sizes for collecting field data with GPS devices by estimation using a modified IET equation; travel speed is described as stable and can be easily measured for travel time and delay studies. Remias, S.et al. (2013) use probe data sources to identify the adaptive control at the intersection. Some very specific problems always exist in the data fusion process related to GPS, for example, the expression of speed and travel time. Li et al. (2014) measure travel time by tracing probe vehicles with GPS, while the average speed data on a link from GPS appears to be closely related to the inverse of the average travel time.

Intersection control plan

Smith, B (2001) describes the three main elements in timing plan: cycle length, splits, and offsets. Nakatsuji, T. et al. (2004) estimate the turning movements at intersections using a logit-based stochastic user equilibrium (SUE) model integrated with a genetic algorithm. Kumar, S. et al. (2011) use only location-based flow data to estimate some spatial parameters such as density and travel time by using LWR model. Banks, J. (2006) investigates the interrelationships between the intervening variables and average flow per lane under capacity conditions.

2.4. Reasoning and positioning

Having went through some of the previous studies, there still exist some. Therefore, this part introduces them from several aspects, and shows reasons for the development of methodology with an intention to face these challenges.

Missing data at a junction

As stated before, there are three kinds of missing data. Some traditional ways, such as historical imputation or simple regression can solve the majority of the cases. Whereas, if there are not enough available direct observations, some more information is required or advanced tools should be developed. However, complex models call for high computation costs and complex calibration.

In the chapter 3 data analysis, there does exist a large scale of flow missing data in the SCATS. In this case, the author tries to start from fast and efficient methods, following a 'simple is good' principle, such as to give a quick imputation of missing values. Fortunately, the target concerned in this article is a junction, where there is a natural comparability among flow observations over time and space. For instance, at urban junctions, traffic from one approaching stream is distributed in parallel lanes; the detectors spread over on each lane also provide values from various locations at the same time. Therefore, the flow data collected at an intersection can be expressed from the time coordinate and the space coordinate (equation 4.1), and missing value is expressed by the similar way (equation 4.2). Using the direct observations from other time or space has become the first and efficient choice (equation 4.3).

Improvements of efficient methods

As stated in the previous part, although many types of research use the imputation both temporally or spatially, seldom of them tells the difference between these two under various situations. To face this challenge, at first, the author implements some primary methods (primary implementation in approach 1) to show the utilities of temporally or spatially imputation, simultaneously and independently. Then the author tries to make improvements based on specific characteristics of these traditional methods. For example, weighting factors are added to improve the reliability of relations between traffic flows observation. Finally, the author makes the integration of these methods by involving the methods in an iteration (integrated method 1).

Traffic flow at signalized junction

A traffic flow q is defined as the number of vehicles passing a certain cross-section within a unit time. Similarly, the traffic flows at the junction can be seen as the traffic

volumes through a lane or several lanes within a given period. Compared to traffic flows on freeways, traffic flows at junctions have some particular characteristics. One of the most important characteristics is the interrupted flow. Taylor et al. (1996) state that, the delay and congestion of uninterrupted flow are generated by internal interactions in the stream of flows themselves; the flow on freeway is an example of uninterrupted flow. While, for interrupted flows, the performances of the streams are also influenced by external factors such as the intersection control or other kinds of modes or even a railway level crossing. The traditional speed-flow relation provides more than one corresponding speed or travel time for a certain flow value. While for interrupted flow, the speed only decline with the increase of the traffic volume. The comparison of this relation is shown in the following figures.

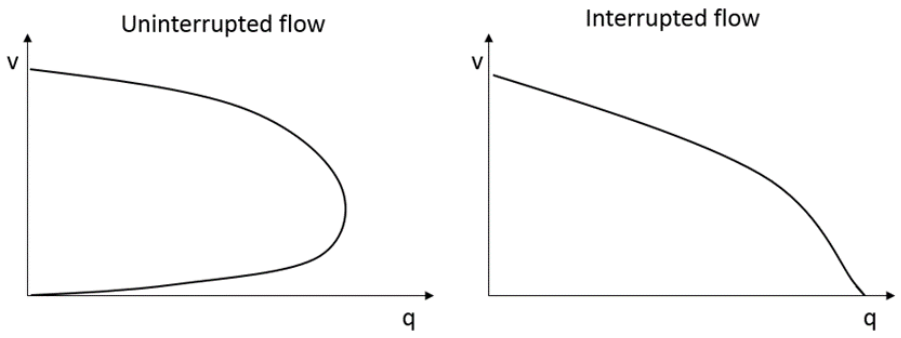


Figure 2-1 Schematic diagram for the relation between uninterrupted flow and interrupted flow, Taylor et al. (1996)

For urban junctions, traffic flow only exists when the green light turns on, that is to say, unlike on freeways, traffic at the urban junction are discretely interrupted by signal lights. Thus, it can be described as a kind of interrupted flow. Therefore, the flow detected at the junction in the thesis is, in fact, a kind of interrupted flow.

These theoretical support have led to a convenience when analyzing the relation between speed and flow at the junction. Although speeds of the corresponding segment are unknown, they are assumed by other ways; for example, FCD helps to represent the speed on the segment (approach 3).

Further in regression

As previous part says, the flow measurement may link to both time and location factor at the same time. However, the underlying relevance between the observation, that is to say, the degree the values can influence and contribute to the missing flow can be rather complex.

The research question, is then inspected from statistical aspect. The author uses regression to train the relations among the observations. As described in the previous parts, the regression is a widely used tool. However, there are some challenges about the regression tool itself, too. Firstly, the specific input type and amount are unknown.

Secondly, the analysis interval of the parameters in the regression model is another difficult consideration.

To face these challenges, The author develops an MLR (multiple linear regression) (approach 4) with various inputs range to see the influence of inputs and, considering the period of the analysis interval from 4 hours to 24 hours. The author also tends to get more understanding of regression by evaluating the performances every one hour, thus to maximize the utilization of the available data (integrated method 2).

Other concerns

Except for the challenges mentioned before, some important concerns in the previous research also play significant roles in the research. The author considers them and tries to make improvements from them. Here are some major examples.

The first concern is about the choice and utilization of relevant days and locations. Except for the neighboring days, observations from other relevant sets of the days are also supposed to be useful. The author provides estimation using different groups of days by adding weighting factors; so as it for the estimation from the spatial dimension. The weighting factors are also derived from the data by finding the relations from data itself, and they are updated with the process of the estimation.

The second concern is about the level of aggregation of the input data. The traffic volumes detected at junction are composed of several traffic during a period, and these outputs from the system has already been aggregated somehow. However, they still hold large deviation of variances over time. The direct observation can be close to ground-truth data after smoothing. In previous studies, some show the results of imputation before aggregation versus aggregation before imputation. In this research work, the characteristics of the data itself are also considered by a different and convenient way: make the estimation according to two types of inputs: using original data (direct observation) and processed data (smoothed).

Another concern is the consideration of other traffic elements. The research work not apply control plan as a reference, the green information from it is also taken into account. For example, Approach 1.2 also uses the concept to try to conduct the relation between flow and green.

A flow chart showing the positioning of the thesis work is presented in Figure 2-2, which also gives the reasoning by using arrows. Specific methods are marked by colored blocks.

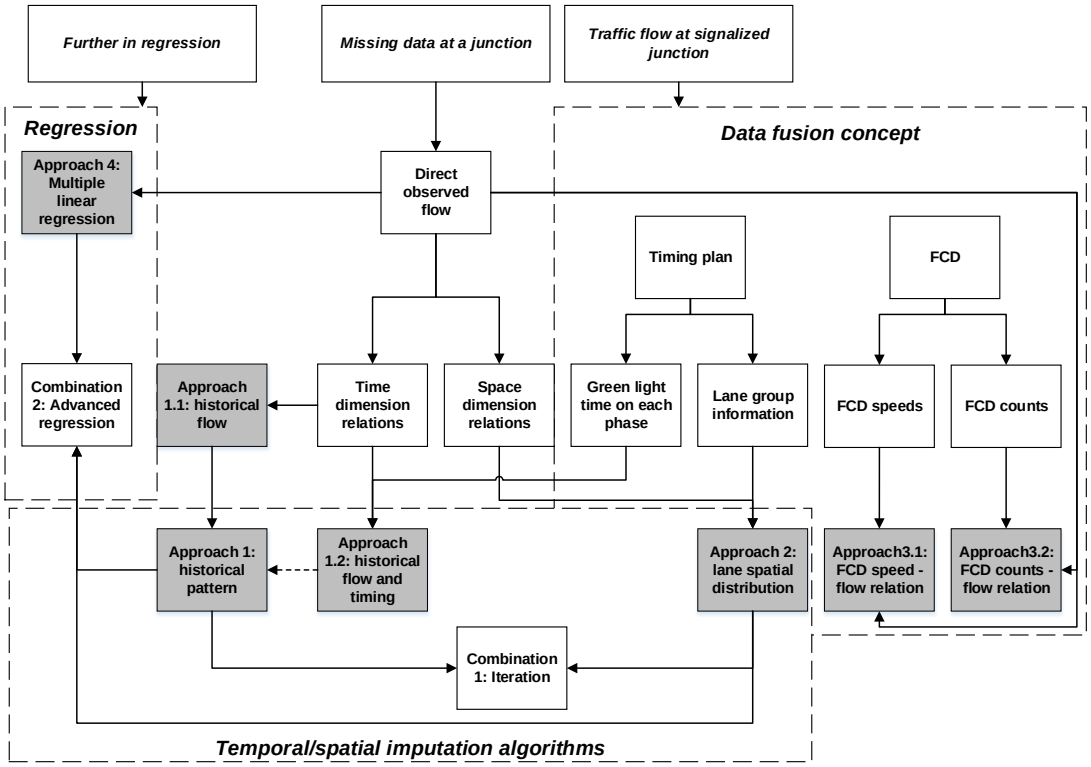


Figure 2-2 flow chart for the position of the work

2.5. Conclusion for the chapter

This chapter has gone through the research related to the topic from the aspect of missing data imputation, data fusion, and others. Major imputation methods are classified. This research not only gives single imputation to replace the missing value, but also makes continuously updates in the estimation. Thus, in the following chapters, the process to get these missing flow will be called as missing flow estimation. The challenging topics according to current methods and situations are raised, followed by the way to face them. These topics are:

- *Simple and efficient methods for fast computation*
- *Targeted at the traffic flow at signalized junction*
- *Dig more out of traditional imputation of temporal and spatial dimension expression*
- *Evaluate and improve the regression methods*

3. Data analysis

Except for the theories, data at hand is an inseparable part to discover the issues and to search for solutions. This chapter specializes in the analysis of the data sources. Three kinds of data sources originated from two independent traffic systems: offline data (flow data, timing plan data) from SCATS and offline FCD (taxi floating car data) from GPS data. Section 3.1 analyzes the quality of loop flow data from SCATS. Section 3.2 presents the patterns of loop flow data. The data type of timing plan and its utilization are described in section 3.3. Finally, Section 3.4 demonstrates the data structure of FCD and possible utilizations. In the end, a conclusion is drawn from this chapter.

3.1. Data quality

The quality of traffic flow data from loop detectors is presented, since they are the main focus of the research. SCATS is implemented to junctions in Changsha city in two terms. 102 junctions are equipped with detectors in the 1st term and 104 junctions are equipped with detectors in the 2nd term.

The data availability is defined as the number of observed data available over the number of observations that should be recorded during a whole day.
Figure 3-1 shows the current data quality in the urban area for one day (23rd April 2013) for junctions in 1st term and 2nd term. The number on X-axis (1 to 102) refers to the junction number. Each blue bar in the figure represents the availability at a junction. The value for every bar ranges from 0 to 1, in which 0 shows that no data is available at this junction, and 1 shows that all detectors work well and provide flow data over each time of the day.

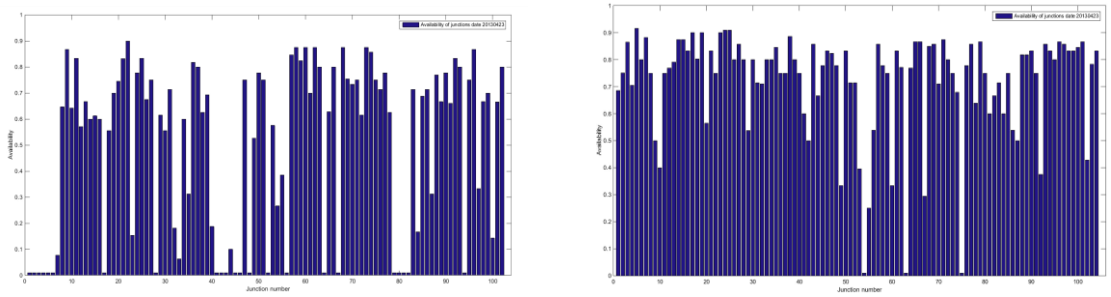


Figure 3-1 Data quality for all 1st term junctions in SCATS (left), data quality for all 2nd term junctions in SCATS (right) (23rd April 2013)

The average rate of data availability is 51% for all the 1st term junctions and 75 % for all the 2nd term junctions in 2013. Looking at the data quality of those junctions distributing over the map, some adjacent junctions are of similar low data quality, and some others are with higher quality. Thus, the data availabilities of these junctions are similar in certain areas. However, this research starts from a certain junction and then expands to multiple junctions level, thus the relations among junctions will be not be considered in this thesis but in future works. Since the junctions from the 2nd term generally have better data availability than the ones in the 1st term, test cases are picked up from the 2nd term junctions due to relatively complete observations. The developed methods are expected to be applied to each single junction from either 1st term or the 2nd term.

According to the analysis of available SCATS data, there are a lot of types of missing flow. For experimental purposes a classification of missing data is defined: The first type is the long-term missing, which lasts for a whole day. The second type of missing data is incidental (random) missing, which only occur at a very short period.

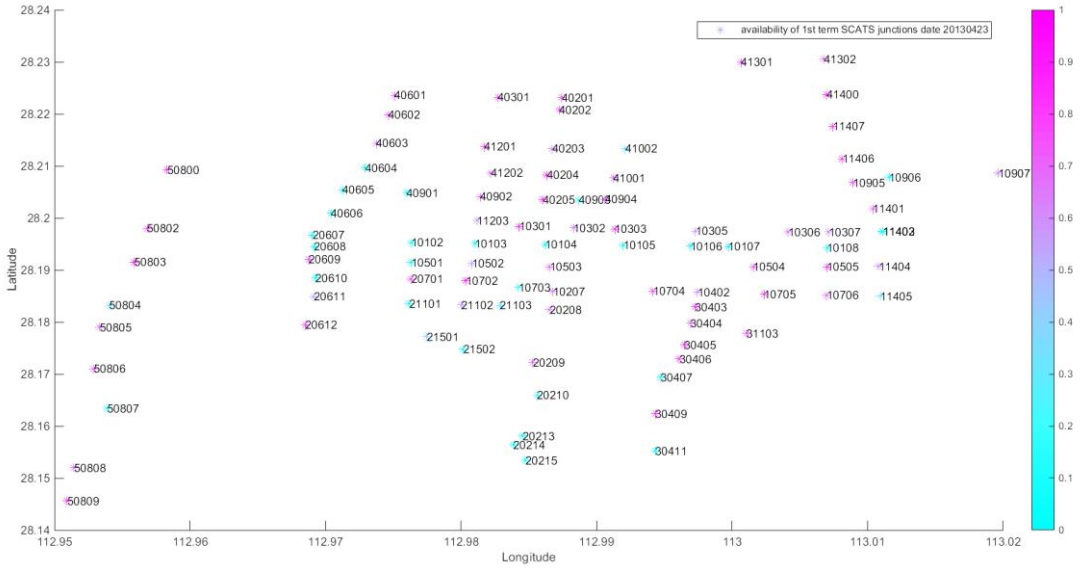


Figure 3-2 Data quality distribution over whole network in Changsha city for all 1st term junctions in SCATS system (23rd April 2013)

3.2. Loop flows

Loop flow observations used in the thesis come from SCATS in Changsha, China. In the figure, the flow pattern over a week on all lanes are visualized. On the one hand, for a single lane, the flows have their similarities in every day of a week; on the other hand, in one same day, the flows have their similarities distributing over the lanes, and these similarities are highly related to the signal plan.

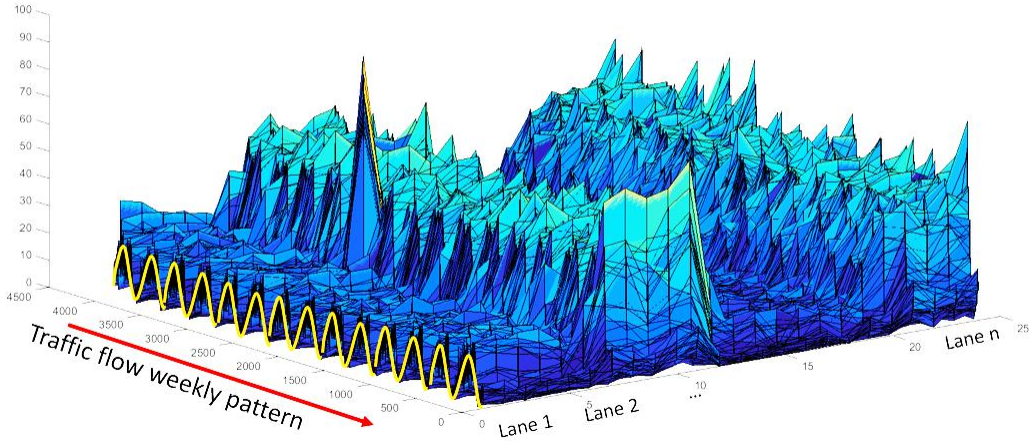


Figure 3-3 Visualization of all the flows for two weeks on all detectors at junction 31616 in SCATS

For a junction, if compressing the cubic onto one surface (day axis), means and standard deviations from all the set of flows at a same location are presented as

follows. No matter for the means nor the variances, they show an obvious trend on the time dimension. For a specific lane, the averages of flows on this lane keep stable, while the variances are also stable. If compressing the visualization cubic onto another surface (lane axis), averages of flow volumes of each time for all the lane are plotted as Figure 3-5 . It can be seen, the flow volumes increase and decrease with almost the same range for each lane over a day.

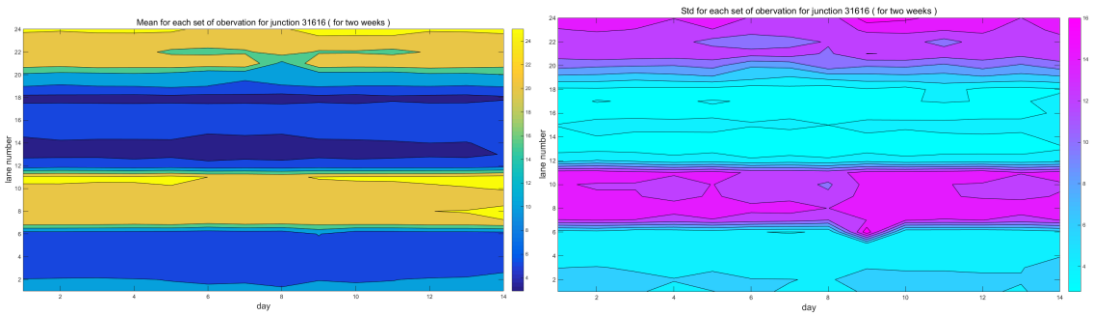


Figure 3-4 flow Mean (left) and standard deviation (right) at junction 31616 in SCATS

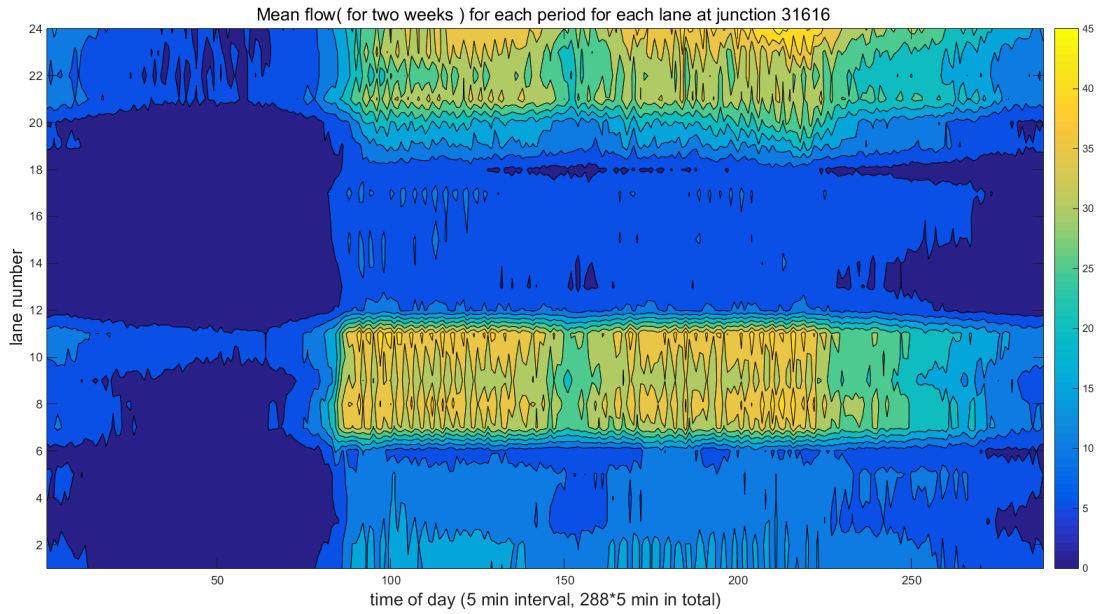


Figure 3-5 Mean traffic volumes for each lane over a day at junction 31616 in SCATS

According to these findings, it is assumed that the flows are correlated both in the time dimension and the spatial dimension.

To verify the assumptions, the correlation coefficient and its P-Value test are used for uncovering the relations. The correlation coefficient illustrates a quantitative measure of some correlation and dependence, and it shows the statistical relationships between two or more random variables or observed data values. Pearson product-moment correlation coefficient method is used. The calculation formula is:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \tag{3.1}$$

Where $cov(X,Y)$ is the covariance and σ_X is the standard deviation of the variable X, σ_Y is the standard deviation of variable Y. To construct a confidence interval around correlation coefficient that has a given probability of containing ρ, the P-value test is conducted. The test shows the probability of an observed result, assuming that the null hypothesis is true, thus here a Hypothesis test is given by:

Null Hypothesis: Ho*:* $\rho = 0$
Alternate Hypothesis: Ha*:* $\rho \neq 0$

Set the significant level as $\alpha = 0.05$. If the P-value is less than the significance level, there is sufficient evidence to conclude that, there is a significant linear relationship between variable X and variable Y. The hypothesis H is rejected if the p-value is less than significant level.

Historical flow pattern

Two individual observations show similar trend, they are on the same location and of the same day of week (DOW) from two weeks. The P value is much smaller than 0.05, thus the results of correlation coefficients are significant.

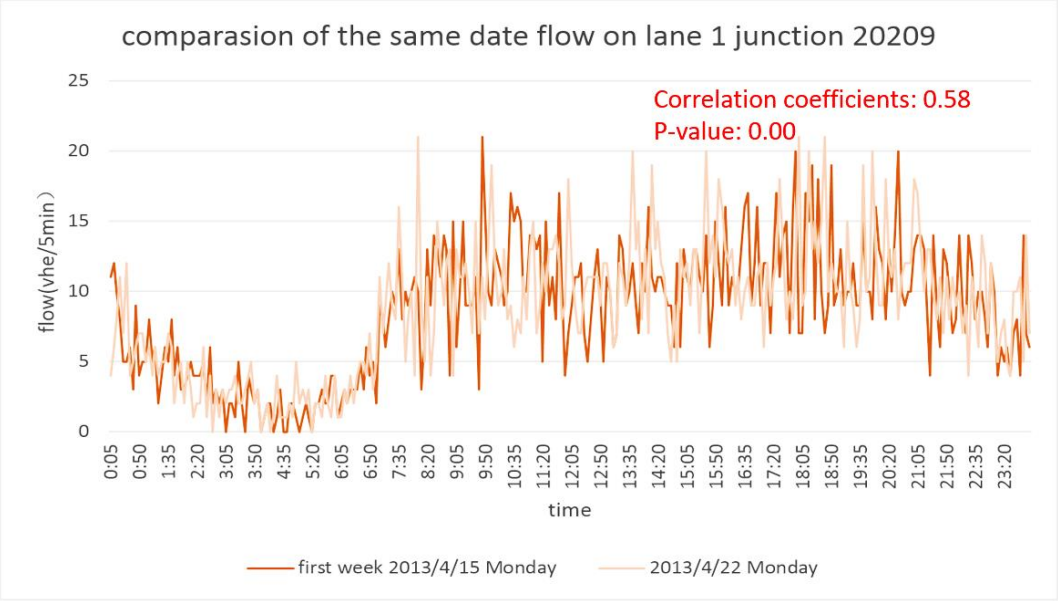


Figure 3-6 Correlation coefficients and p-values (in brackets) from two days with the same DOW

If the sets of observations from all the available days in the same lane are put together, the results become this 'Correlation coefficient map', which showing correlations of the flows from each pair of days. (In the appendix 6 , more figures are aviliable)

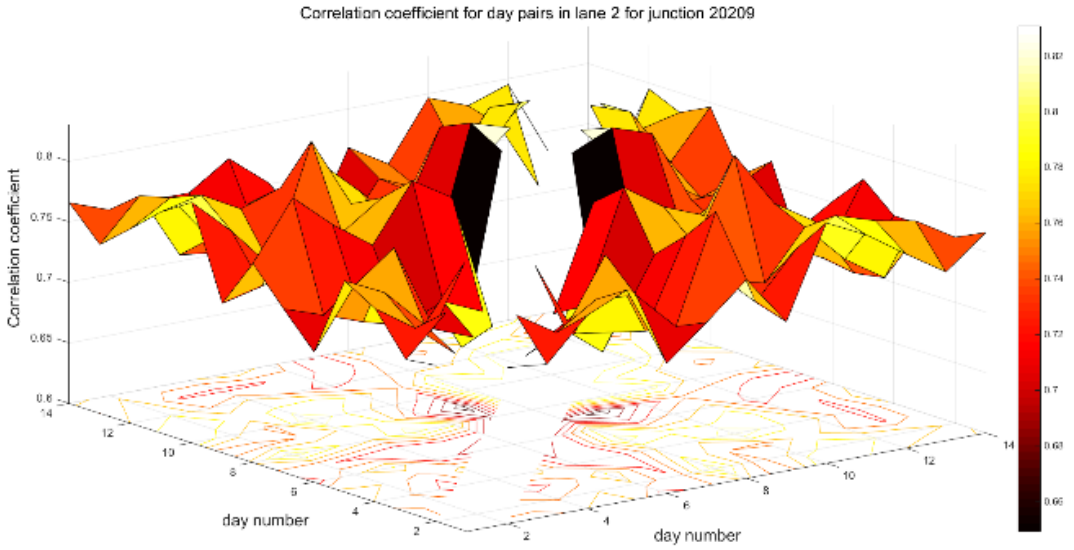
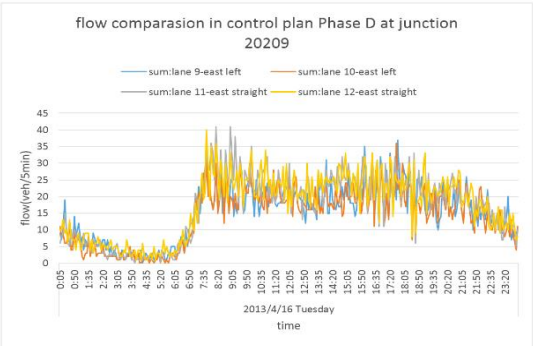


Figure 3-7 Correlation coefficient map for lane 2 at junction 20209

The correlations level of flows from days at the same lane are distributed relatively smooth. The P values are all quite small, thus the direct ratios between the sets of flows are significant. Thus, it is assumed that, for a same lane, the observations from other days are relevant and useful for an estimation.

Flow relations according to lane distribution

Similar to the historical pattern, flows from certain lanes in the same day show similarities. For all pairs of variables, The P values are all smaller than0.05, thus the results are all significant. Similarly, correlation coefficient maps can showall the correlation coefficients of flows onall pairs of lanes in a same day with available data...



	Lane 9	Lane 10	Lane 11	Lane 12
Lane 9	1 (1)	0.87 (0.00)	0.89 (0.00)	0.67 (0.00)
Lane 10	0.87 (0.00)	1 (1) (0.00)	0.92 (0.00)	0.64 (0.00)
Lane 11	0.89 (0.00)	0.92 (0.00)	1 (1)	0.66 (0.00)
Lane 12	0.67 (0.00)	0.64 (0.00)	0.66(0.00)	1 (1)

Figure 3-8 Correlation coefficients and its p-values (in brackets) from four lanes in a same phase

On the correlation coefficient map, some peaks are observed, for example, the correlations among lane 12, 13 and 14 are extremely high. They are exactly in the same

phase. A detailed contour compare the correlations with the control plan for junction 31616. It is obvious that the lanes in the same control cycle group (phase) have higher correlation coefficients. This relation can be also observed on other junctions. For example, correlation coefficient maps from junction 20209 and junction 31617 are presented in appendix 6.

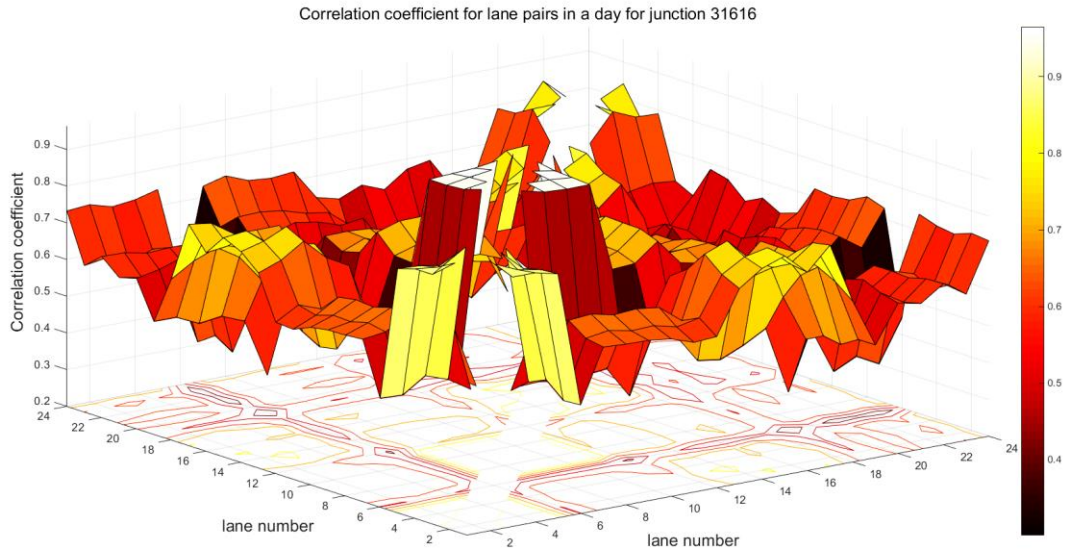


Figure 3-9 Correlation coefficients map between lanes for junction 31616 in SCATS.

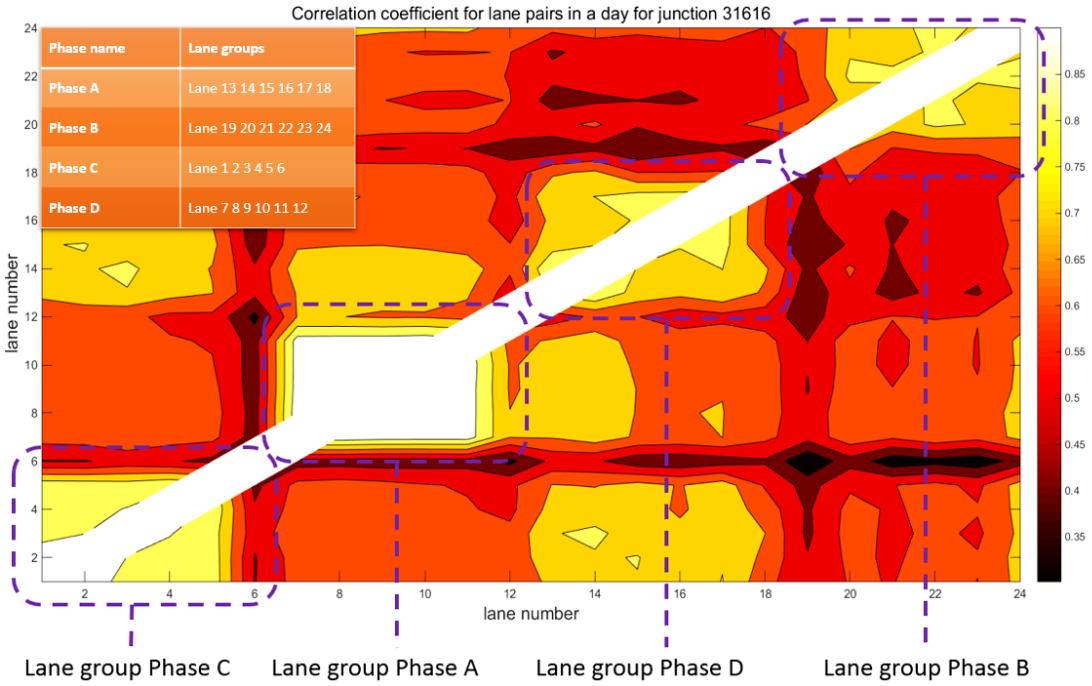


Figure 3-10 Correlation coefficients contour figure between lanes for junction 31616 in SCATS.

Except for the lanes in the same phase, the lanes from the same stream that close to each other also have a high correlations. Although not shown here, the turning is also

a large factor for the distribution of lane flows. The actual mechanism of how flow influenced by the lane distribution may fall in all of these reasons simultaneously.

To make a conclusion, the lanes in relative groups are holding high correlation coefficients, which indicates these lanes are relevant with each other. The statement is going to be a strong support of missing flow estimation approach 2- Lane spatial distribution.

3.3. Timing plan

Apart from the loop flows observations, signal timing plan is another data source from the same system (SCATS) but with a different type. The availability of signal timing plan data is optimistic after checking all the samples: almost all the timing plan data recorded are normal. From timing plan, two kinds of information can be obtained: 1. the lane groups in the same phase. 2. Duration of each phase and starting time of a cycle (the accuracy is up to one minute).

The groups of lanes in the same phase have been used in the analysis of flow relation in the previous part. For this part, the duration of a phase is introduced. The duration of a phase shows the green and red a cycle gives to a stream. Since a traffic stream can only pass the junction under green light, it is assumed that the traffic volume and green light time should be somehow related. To make a connection between the timing plan and traffic flows, a timing plan- flow diagram is drawn. In this diagram, the horizontal axis shows the time and the vertical axis shows the distribution of lanes. Several lanes are grouped in a same single phase. Green rectangles represent green light amount in a phase. White numbers are the traffic volumes from each detection interval.

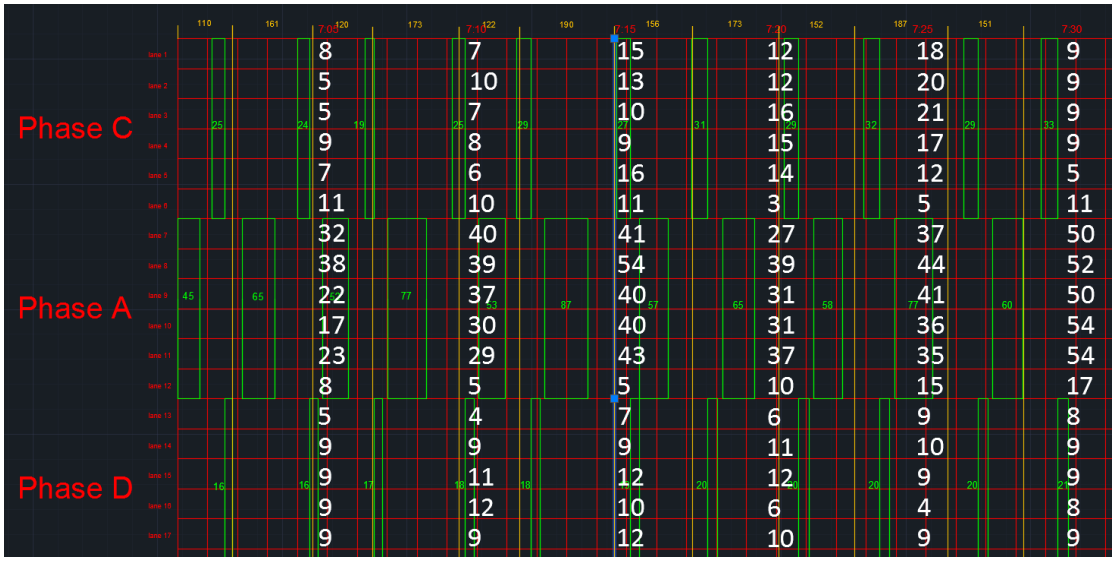


Figure 3-11 Sample of timing expression, flow values are involved.

The data from the diagram are manipulated to get more information: From the

available data, only the starting time stamp of each control cycle and the green light time of each phase of a cycle are known. The starting time stamp is in integer minutes such as 7:05 and 7:07. Therefore, the precise duration of a cycle is unknown since there lacking information of yellow lights and all-red lights time. However, the total time of these yellow lights and all-red lights can be assumed by the comparing the starting time of two cycles with a certain distance with the adding up green light time of all cycles in between. In this case, for a total 30-minute interval from 7:00 to 7:30, 11 cycles are exactly involved. A match between the control cycles and flow detection interval can be set by finding a period with their least common multiple. Similarly, the green lights in cycle and the flows in each detection interval can also be matched. By synchronizing lane flow with timing plan, it helps to involve the timing plan to flow estimation.

To make a conclusion, the timing plan data can contribute to the flow estimation by two means:

- *Grouping the lanes in a same phase, which is useful for lane spatial distribution (approach 2).*
- *Matching the phase with flow observations (approach 1.1) and FCD (approach 3).*

3.4. FCD

The raw data of FCD are from Taxi GPS, in Changsha, China. As the Figure 3-12 shows, there are several important attributes that could be directly or indirectly used to form trajectories for a taxi. Here gives an introduction of them and relative usage of them.

- *Vehicle (taxi) ID*
- *Time of day*
- *Recording day*
- *Instantaneous speed*
- *Latitude and longitude*
- *The Instantaneous heading*

Taxi ID defines a certain vehicle. Time of day distinguishes the days of records. By sorting ID and day, a general route of a certain vehicle can be presented.

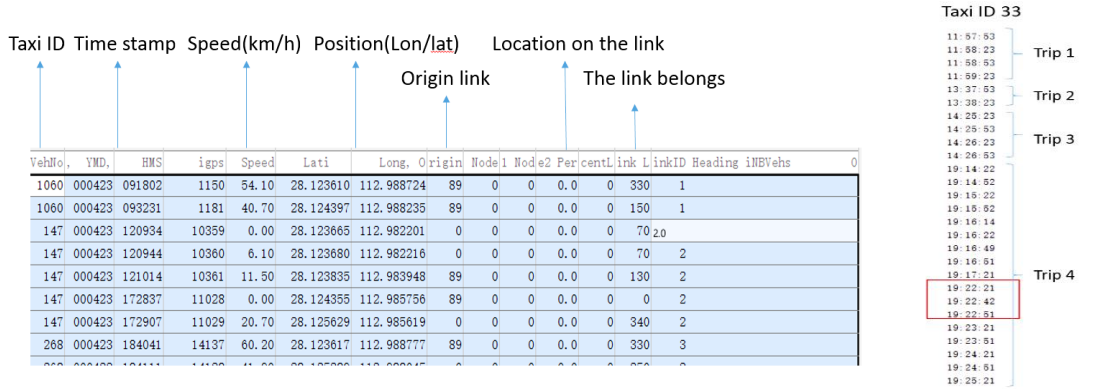
The time of day (timestamp) of a taxi can imply the trips during a day. Regularly, for a certain ID, there is a 30-second interval between two consequence records. However, there are also some exceptions, taxi ID 33, it has 21-second intervals here. Even for a same taxi, running at a place during different period are seen as different trips. For example, taxi 33, has 4 trips according to its timestamp differentiation.

Speed is the instance velocity of a taxi at that timestamp. It can represent the velocity

vector of a taxi combined with heading. The unit is km/h in the raw data, and it should be transferred to m/s when needed. In the areas near junctions, nearly half of the speed records are nearly zero. This represents, the queuing status at the signal controlled junction.

Longitude and Latitude show the position of the taxi at a certain time. A general scatter plot gives the geographic distribution of taxis at a certain time (Figure 3-14). Besides, these two attributes don't provide distance directly; more processes are need for the transformation to distance information. (See section 5.2)

Original link and link ID are for the link match of the taxi. However, there lacks a link ID map; thus link match could only be carried out manually by self-made methods. The methods are shown in the data processing chapter.



VehNo	YMD	HMS	igps	Speed	Lati	Long	Origin	Node1	Node2	Per	cent	Link L	linkID	Heading	iNBVehs	0
1060	000423	091802	1150	54.10	28.123610	112.988724	89	0	0	0.0	0	330	1			
1060	000423	093231	1181	40.70	28.124397	112.988235	89	0	0	0.0	0	150	1			
147	000423	120934	10359	0.00	28.123665	112.982201	0	0	0	0.0	0	70	2.0			
147	000423	120944	10360	6.10	28.123680	112.982216	0	0	0	0.0	0	70	2			
147	000423	121014	10361	11.50	28.123835	112.983948	89	0	0	0.0	0	130	2			
147	000423	172837	11028	0.00	28.124355	112.985756	89	0	0	0.0	0	0	2			
147	000423	172907	11029	20.70	28.125629	112.985619	0	0	0	0.0	0	340	2			
268	000423	184041	14137	60.20	28.123617	112.988777	89	0	0	0.0	0	330	3			

Figure 3-12 Example of data format of FCD and trip determination

The data availabilities for Vehicle ID, Latitude, and Longitude, Time of a day, the general availability are ideal. While, for instantaneous speed and heading, their availabilities are low in some records. Besides, Latitude and longitude are not matched with a map; this has been revised by checking and switching the global deviation.

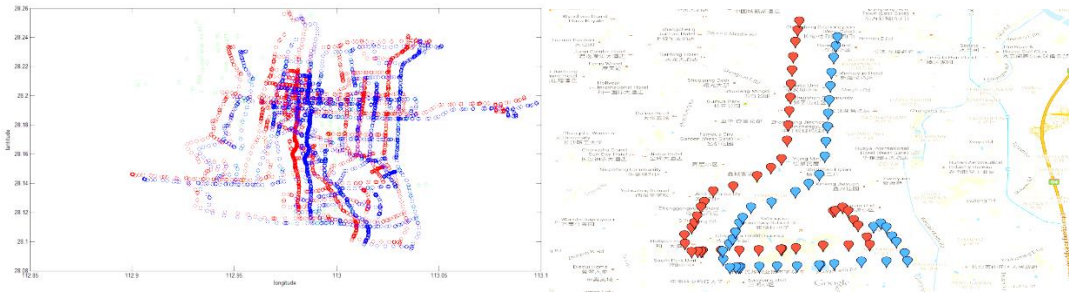


Figure 3-13 an example of a map-matching check using Google map

When the speeds of floating cars are demonstrated on the network by different colors, some traffic status can be visualized. (Here in the figure, the green color represents free speed, such as speed over 60 km/h. While red color represents low speed from 10 to 20km/h, and black shows very low speed which is lower than 10km/h. Other colors stand for the speeds in between.)

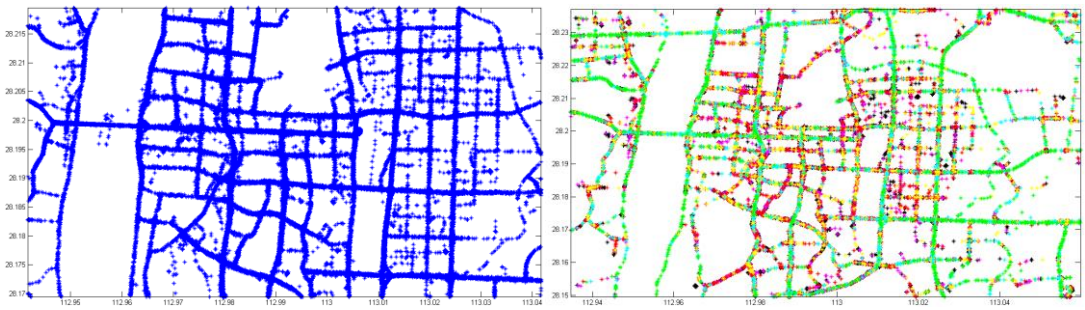


Figure 3-14 FCD plots using one tenth of all the taxi (left) and speed plot (right) 23th April 2013

From the figure, it is clear that the majority of the low-speed records are located at junctions. What is more, the records of the taxis are comply with the shapes of the road segments. Therefore, matching the FCD to each link and junction segment is possible.

Two kinds of information from FCD can be expected after data processing, they are FCD speed and FCD counts. Both of them come out from FCD trajectories. A specific way to process FCD trajectories will be introduced in the data processing chapter. In this chapter, only the results from the analysis are presented.

FCD speed

By visualizing FCD, the speeds during each period are shown. Data in an area is selected to give an example. As the Figure 3-15shows, the research area is the road between junction Lao dong road/Fu Rong road and Lao dong road/Shao Shan, which is a west-east road. The SCATS number of Lao dong road/Fu Rong road is 20209, and the SCATS number of Lao dong road/Shao Shan road is 30407. The period is 6:45 to 8:45 in the morning of 2013.4.23 and the time interval is 15min. On the speed plot, red represents higher speed and blue represents lower speed.

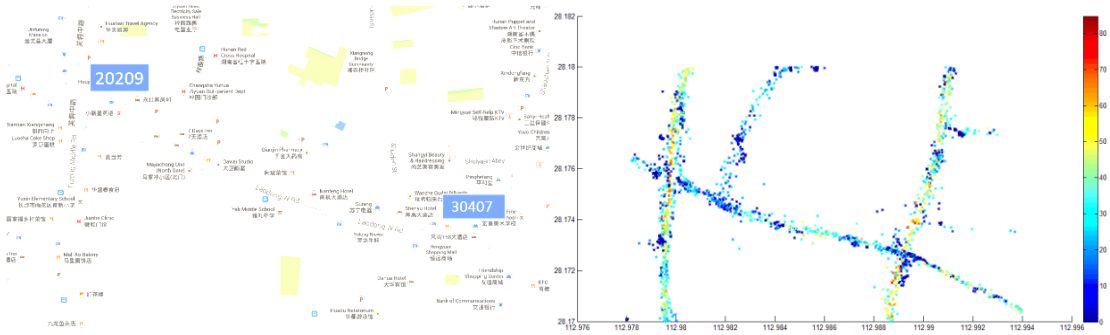


Figure 3-15 Layout of research areas and FCD records showing speed on Lao Dong road from 7:00 to 7:15 23th April 2013

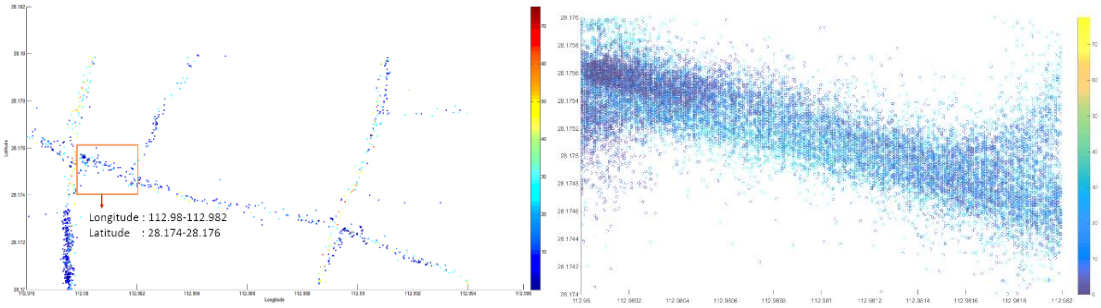


Figure 3-16 Chosen area (Longitude: 112.98-112.982 Latitude: 28.174-28.176) and FCD vehicle speed during a day in the small area near junction 20209, 23th April 2013.

Here is an example of the speed plot near junction. To make a comparison for the possible data fusion process, an area near the junction is chosen for analysis. FCD can be processed to provide reference speed at this road segment. Besides, the counts, together with the flow counts at the same position, can give penetration rate of the taxi during this period.

FCD counts

The FCD counts are collected by processing raw data by vehicle ID, position and time. These FCD counts are compared to the traffic flow detected by the nearest neighboring loop, is the ratios are seen as the penetration rate at this location.

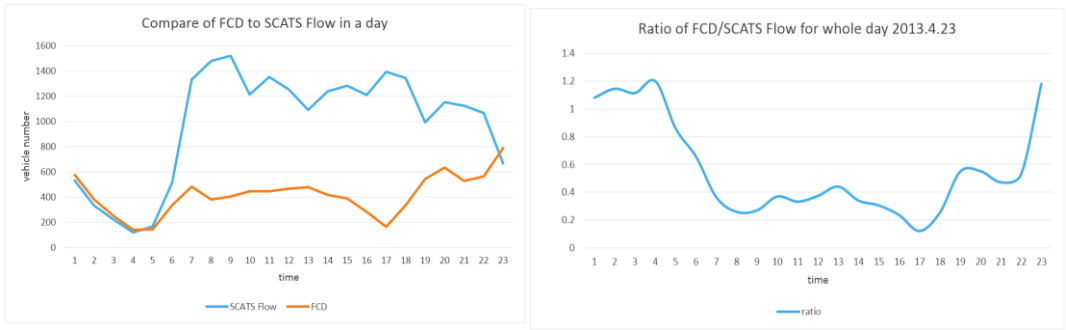


Figure 3-17 Compare of FCD/Flow in SCATS (left) and the ratio (right) at Lao Dong road between 0:00-24:00 23th April 2013

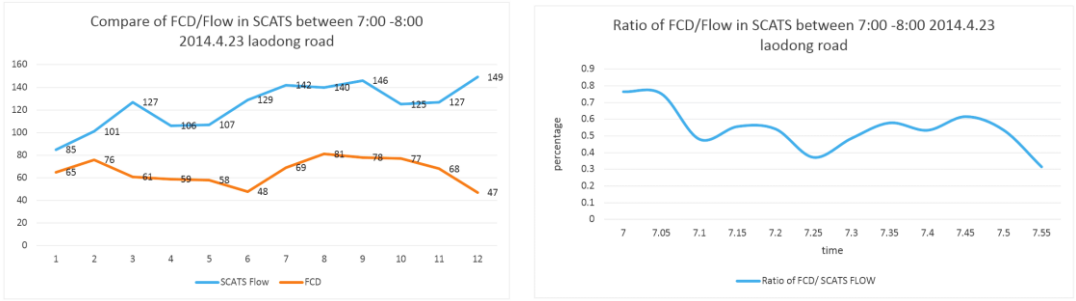


Figure 3-18 Compare of FCD/Flow in SCATS (left) and the ratio (right) at Lao Dong road between 7:00 - 8:00 23th April 2013

The penetration rate location is low during the morning peak and afternoon peak,

however, it becomes quite high at midnight. Rooming into the peak hours, the ratio has reached 0.7 at 7:00. This due to that, many taxis are still on road during the mid-night, but social vehicles increase largely during the daytime and increase sharply during the night. Besides, in this case, the location is at the city center, which is a highly utilized link for route choice for taxis.

3.5. Conclusion for the chapter

This part tells important facts that, nearly one-third (49% in 1st term and 25% in 2nd term) of the flow data are missing from SCATS. Both long-term and short-term flow missing cases exist. From the analysis of flow means and variances, similarity patterns obviously exist in flow observations over time and space. Besides, these relations are proved to be direct correlations with high correlation coefficients. This chapter also links timing plan to the flow observations, showing that lanes are grouped by phases and that flows can be matched with timing. For FCD, the author has shown the desired coverage on the urban network. Developed from the attributes, two utilizations are confirmed: speed and counts. In conclusion, this part has shown the relations within and between data, which are stable and reliable supports for the next step analysis and algorithm developments.

4. Methodology

Chapter 4 provides the methodology, including four individual approaches and two integrated methods. The four individual approaches are from two aspects: estimate a missing flow by using other directly flow observations, or by assuming a traffic flow theory and using a data fusion concept. Firstly, to ensure a unification, section 4.1 raises the general framework of expressions. Secondly, section 4.2 provides individual approaches: historical pattern (including two sub-approaches), lane spatial distribution, FCD and traffic flow data fusion (including two sub-approaches), and MLR (Multiple linear regression). Thirdly, section 4.3 shows two integrations. The first combines historical pattern and lane spatial distribution methods using an iterative process. The second improves MLR by referring information from the historical pattern and lane spatial distribution. Finally, sections 4.4 makes a conclusion of this chapter.

4.1. General framework

In the beginning of the descriptions of approaches, the flows and missing flows are expressed by the way they are measured. Then, all the approaches and methods are expressed.

Expressions of flow

Detectors provide flows in a certain interval at a specific day, and usually, these detectors are fixed at a location. Therefore, a flow observation, according to the way it is measured (measured by detectors), is defined by location, time, and TOD (time of day).

$$q(l,d,t), l \in L, d \in D, t \in T \tag{4.1}$$

q: *Estimated traffic flow;*
l: *Location of an observation where it is detected;*
d: *Day of an observation when it is detected;*
t: *The time of an observation when it is detected;*
L: *Set of lanes concerned;*
D: *Set of days concerned;*
T: *Period of time concerned;*

Similarly, a missing flow can be expressed by adding a subscript x to the variables.

$$q(l_x,d_x,t_x), l_x \in L, d_x \in D, t_x \in T_x \tag{4.2}$$

l_x: *Location where the missing data located;*
d_x: *The day when the missing data located;*
t_x: *Specific time of the missing data;*
T_x: *Set of period of time related to the missing flow time*

- ***Estimation from direct observation***

Naturally, when the flow observation is missing from a detector, other observations are considered. So the first choice is to use the direct observations with different location, day, or time. Since the flows concerned are detected near the urban junctions, usually signal timing plan can also help:

$$\hat{q}(l_x,d_x,t_x) \sim \{f(q(l,d,t)|l \in L, d \in D, t \in T), \{s(L_x,d,T_x)|d \in D\}\} \tag{4.3}$$

f: *Function to operate flow observations;*
s: *Signal timing plan;*
L_x:*The group of lanes that containing the lane with missing flow;*
T_x:*The period that containing the period with missing flow;*

Some representative approaches using this way are Approaches 1, 2, and 4.

- ***Estimation from data fusion concept***

The second choice is to use measurements from other independent data sources. One possible way is refer to the traffic flow theory such as $q = k * v$.

$$\hat{q}(l_x, d_x, t_x) \sim \{q_{ext}(L_x, d_x, T_x), u_{ext}(L_x, d_x, T_x), k_{ext}(L_x, d_x, T_x) .. etc\} \quad (4.4)$$

u: *Average vehicle speed;* k: *Density, the number of vehicles per unit length of the road;*
ext: *External /other data sources*
g: *Function describing relation of flow to density/speed;*

Suitable datasets can be many: FCD (speed, counts), CCTV camera (flow, speed, and travel time), AVI etc. The Approach 3 is one example of this kind of estimation.

Decomposition of the framework

The general framework considers the issue from the aspects of direct observation and data fusion. In this part, each part of the general framework (4.3) (4.4) will be introduced separately.

- ***Part 1: The available direct traffic flow observation***

The first part shows direct observed traffic flows from another time and space. The inputs of this part are sets of flows with various coordinates. The formulation is as follows.

$$f(q(l, d, t) | l \in L, d \in D, t \in T) \quad (4.5)$$

To ensure there is enough relevance between the inputs and the missing values, at least one of the coordinates (l, d *or* t) should be fixed to where or when the missing values are. For example, the sets of days can be used only when it is at a same lane. Similarly, it only makes sense to compare the flows on different lanes on a same day. However, there are two situations that may ignore this restriction. Firstly, all the values are retrieved iteratively and cross-compared. This situation will be introduced in the first integrated methods. Secondly, if the function f is a regression tool (which will be later introduced as the approach 4 and integrated method 2), the sets of input values are given more freedom, this is because that, regression tool has the ability to distinguish the relevance by itself.

Usually, the coordinate time of day t is fixed, that is to say, values are chosen from other days or other lanes but the same time period in a day. There is one exception, for a detection failure that lasts only for a short period of time, the available values shortly before and after this period are also importance. In this case, it is suggested to take

into consideration the flow values over nearby periods in the same day on a same lane.

Another important component is the function f, which may be composed of sequences of algorithms or processes, indicating how the selected values are calculated, assimilated, or normalized towards the missing values. Generally, there are many choices for those algorithms or tools.

- ***Part 2: The signal timing (control) plan***

This part shows signal timing (control) plan. It plays three main roles in the estimation. Firstly, the group information from control plan is used to select lanes on the spatial dimension; which will be introduced in Approach 2. Secondly, green light time is assumed to be related with the flows. Thirdly, control plan acts as a reference for of FCD trajectories. The sets of signal timing plan information are denoted as:

$$s(L', d, T'), L' \in L, d \in D, T' \in T \tag{4.6}$$

For time coordinates, the day is the same day d when the detection fail. As for the specific time of a day, a signal timing plan is not as the same interval as flows does. Here T' shows a period containing one or several cycles, covering the time during which the flows are missing. For the spatial coordinate, several batches of lanes are grouped (maybe in a same phase). L' stands for the group of lanes that containing the lane with missing flow.

- ***Part 3: The external sources part***

This part shows the consideration of external data sources. It expresses the effort to make an estimation using data fusion concept.

$$q_{ext}(L_x, d_x, T_x), u_{ext}(L_x, d_x, T_x), k_{ext}(L_x, d_x, T_x) .. etc \tag{4.7}$$

Some traffic measurements are from other ways, with different formats and resolutions. The data fusion can integrate multiple data into a consistent, accurate, and useful representation. In this thesis work, FCD is an external source to help to make the estimation of flows.

To sum up, considering the available data sources in this thesis, the general framework for the estimation can be expressed in following diagram:

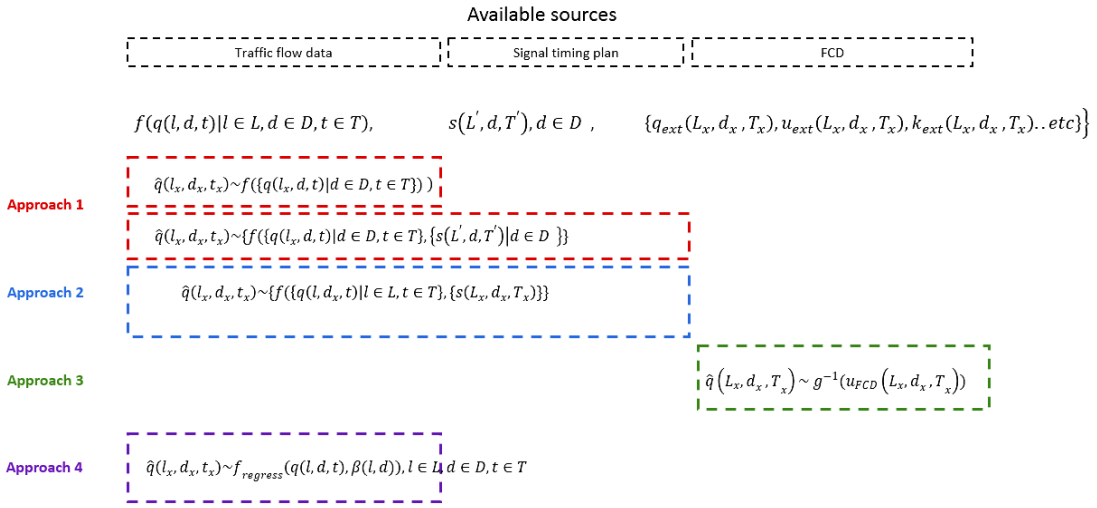


Figure 4-1 the framework decomposition, sources and corresponding approaches

In this diagram, horizontal blocks show the available data sources, and vertical blocks show four individual approaches. Their corresponding formulations are represented in the same color, covering horizontally the sources have been used. In the next part, these individual approaches are introduced in detail.

4.2. Individual approaches

Derived from the main framework, four individual approaches are shown, they considers the issue from the aspects of: (1) historical pattern (both flow and timing), (2) lane spatial distribution (in a timing plan), (3) FCD speed- flow relation, FCD counts-flow relation (4) MLR (Multiple linear regression). Followed by integrated methods (1) Integration using iteration and (2) Improved MLR.

4.2.1. Approach 1 Historical pattern

In the Approach 1 : historical pattern, the historical information of flow are considered independently, and then the flows and timing are considered together as an historical information.

4.2.1.1. Historical flow pattern (Approach 1.1)

The algorithms in this approach retrieve exist observations and use them to estimate the missing flows. The observations are from the time dimension at the same location (on the same lane). The process then keeps updating the multiple missing flow values by finding the balance between each set of patterns from each set of days.

Formulation

$$\hat{q}(l_x, d_x, t_x) \sim \{w_{DOW} * \bar{q}(l_x, d_{DOW_x}, t), w_{WD} * \bar{q}(l_x, d_{WD_x}, t), w_D * \bar{q}(l_x, d_{AD}, t), t \in T_x\} \quad (4.8)$$

Q: *Estimated traffic flow on an aggregate level;* s: *Signal timing plan*
l: *Location of a traffic state where it is detected;* l_x: *Location where the missing data located;*
d: *Day of a traffic state when it is detected;* d_x: *The day when the missing data located;*
r: *Aggregate time of 30 minutes;* D: *Set of days*
AD: *Set of all the days that are available in a relatively long-term*
WD: *Set of days in a same week*
DOW: *Set of same day of week over a sequence of weeks*

Considering different sets of the day, the formula can be also stated as:

$$\hat{q}(l_x, d_x, t_x) \sim \bar{q}(l_x, d, t), d \in DOW_x, t \in T_x \quad (4.9)$$

$$\hat{q}(l_x, d_x, t_x) \sim \bar{q}(l_x, d, t), d \in WD_x, t \in T_x \quad (4.10)$$

$$\hat{q}(l_x, d_x, t_x) \sim \bar{q}(l_x, d, t), d \in AD, t \in T_x \quad (4.11)$$

They represent the estimation only considers the flow from a same week/ the same day of week/all the recent days. The following paragraphs will then describe the development of this approach.

Development

From the general framework, this approach starts from the direct observations in the time dimension:

$$\hat{q}(l_x, d_x, t_x) \sim f(\{q(l_x, d, t) | d \in D, t \in T\}) \quad (4.12)$$

In many cases, the coordinate time of day t is fixed, the main task in this approach is to determine the sets of days. The sets of days are denoted as D and the selection of them depends on: (1) Availability (2) Relevance (3) Reliability. The following paragraphs describe these three points in detail.

- ***Availability***

The availability shows whether a data is recorded at a certain point. To check it, fix the position coordinate l to l_x where missing flow occurs (at the urban intersection, this location is usually a specific lane), search on the time dimension, and update the sets of D. For another time dimension variable t, the availability depends on the type of missing data. For large scale missing type, there is no available T onwards and afterwards of t_x in the same day d_x, so available T onwards

and afterwards of t_x have to be found in other days D. For small scale missing type, additionally, a within-day period T_x onwards and afterwards of t_x is also available.

- ***Relevance***

The relevance shows whether the coordinates are relevant or representative to the missing values. The concept is similar to time-series theory, which is used widely in freeways traffic flow estimation. Four major components in a time-series process, they are Trend, Seasonality, Cyclic and Randomness. The relevant of coordinates are selected referring to these four considerations. Firstly, the longer-term trend is captured by selecting all the available recent days (AD) that are approaching the missing flow day. For example, all the days with available data at one location in one month can be selected a set of AD. Secondly, due to the reason that there is no more detailed information about the categories of flows, the seasonality cannot be identified. Thirdly, for cyclic, the set of days in a same week (WD) and the set of same day of the week over a sequence of weeks (DOW) are used. They show the pattern that are repeated over relatively short spans. (E.g. days in a same week (WD) refers to the last week of January, days in a same day of week (DOW) refers each Monday of January). Lastly, randomness refers to short-term variations. There has been many research and tools looking insight into this aspect. However, this thesis work focuses on deterministic part of the estimation. The probabilistic part will be added to the research work in the future.

Initial choice of D considers the days that are relevant. There are several ways of choosing specific days. For example, there is a way for an ex-post analysis:

- *AD: all the recent available days in a month.*
- *WD: the other days from a same week*
- *DOW: the previous DOWs (day of week) in a month*
- *TOD: several minutes before and after the missing flow (missing data type: small scale)*

For an ex-ante analysis, the situation has changed. In reality, especially for an on-line control approach, there is no access to the information in the upcoming days. Therefore, only data from the past can be used for these sets.

- ***Reliability***

The concern of reliability sorts the values that are already in the sets. For all the sets of D, they are made with an initial choice (already conducted in the relevance part) followed by the updating process (in this reliability part).

By using the initial sets of days D, some estimation results can be obtained, these results are called a temporal reference. Compare the temporal reference with inputs, if the differences have exceeded an unacceptable level, the inputs are considered as

unreliable. As a result, the concerned days will be removed from the sets of days D. The final estimated values come out progressively. The acceptable level is determined by: Assume all the reference values to be of a normal distribution. The expectation of the mean is represented by the temporal reference, and the variance is the same as it in original data. Set 99.7% as the range of reliable confidence interval, so the values within $\mu \pm 3\sigma$ are supposed to be reliable. Finally, update the estimated values until reaching convergence.

4.2.1.2. Historical timing + flow pattern (Approach 1.2)

This approach links flows and timing plan, and takes the flow/green ratio (historical passing ratio) as a new variable to be considered. Firstly, the ratios during each period on each day are calculated. Secondly, an estimation of the ratios on a specific day are made by referring to historical ratios (in the same way as approach 1.1). Finally, the flows are estimated by multiplying the estimated ratios with given green.

Formulation

$$\hat{q}(l_x, d_x, t_x) \sim s(l_x, d_x, T_x) * \{w_{DOW} * \bar{r}\left(l_x, d_{DOW_x}, t\right), w_{WD} * \bar{r}\left(l_x, d_{WD_x}, t\right), w_D * \bar{r}(l_x, d_{AD}, t), t \in T_x\} \quad (4.13)$$

$\hat{q}$: *Estimated traffic flow;*
l_x: *Location where the missing data located;*
d_x: *Day when the missing data located;*
t_x: *Specific time of the missing data;*
T_x: *Set of period of time related to the missing flow time;*
s: *Signal green light;*
l_x: *Location where the missing data located;*
r: *Ratio of flow/green light (vehicle per unit green time);*

Development

The formula involves both flow and green light time.

$$\hat{q}(l_x, d_x, t_x) \sim \{f(\{q(l_x, d, t) | d \in D, t \in T\}, \{s\left(L', d, T'\right) \big| d \in D\} \quad (4.14)$$

The greens in each phase are denoted as $s(l_x, d, T_x)$. The traffic flows on each lane are summed up in the same period $q(l_x, d, T_x)$. The ratio is defined as the flow per green light time.

$$r = q(l_x, d, T_x)/s(l_x, d, T_x) \quad (4.15)$$

4.2.2. Approach 2: Lane spatial distribution

This approach considers the issue from spatial dimension. The first step is to get the estimated value from another individual lane $q_l(l_x, d_x, t_x)$, and then these values are

weighted according to their spatial characteristics to reach a final estimation.

Formulation

$$\hat{q}_l(l_x, d_x, t_x) \sim \left(\frac{\bar{q}(l_x, d_x, t)}{\bar{q}(l, d_x, t)}, t \in T\right) * (\bar{q}(l, d_x, t), t \in T_x) \tag{4.16}$$

$$\hat{q}(l_x, d_x, t_x) \sim \sum_{l \epsilon L} w_c(l) * w_{tr}(l) * w_{ph}(l) * \hat{q}_l(l_x, d_x, t_x)$$

$$\sum_{l \in L} w_{ph}(l) = 1 \sum_{l \in L} w_c(l) = 1, \sum_{l \in L} w_{tr}(l) = 1$$

w_{pl}*: Weighting factors from the aspect of the lane in a same phase*
w_c*: Weighting factors from the aspect of the closeness of position of lanes*
w_{tr}*: Weighting factors from the aspect of the turning similarity of lanes*

Development

The formula is expressed as:

$$\hat{q}(l_x, d_x, t_x) \sim \{f(\{q(l, d_x, t) | l \in L, t \in T\}, \{s(L_x, d_x, T_x)\}\} \tag{4.17}$$

Similar as approach 1, this approach also considers the data in three aspects: availability, relevance and reliability.

Differently, on the time dimension, coordinate day d has been fixed; lane l, from spatial dimension, has been a coordinate to retrieve data. The considerations of availability and reliability are almost the same. The relevance of other location coordinates to the location coordinate with missing value is introduced here:

Firstly, from data analysis, the flows distributed on lanes in a same phase have turned out high correlations. Secondly, turnings of the lanes are always a factor in the research of intersection. Thirdly, lanes that are close to each other also show similarities. Therefore, there are several ways of selection of the sets of L:

- *Lanes in a same timing plan phase, by the phase group (PL)*
- *Lanes that have the same turning, by the group of turning within the phase group (TL)*
- *Lanes that are close to each other, by the closeness of position (CL)*

Although inclusion or overlap may exist among these sets, such as: $PL, TL, CL \subseteq L$ *or* $TL \subseteq PL, CL \subseteq PL, CL \subseteq TL$.

Different from the historical pattern approach, this approach makes the estimation from each individual lane and then combines them. Let $q_l(l_x, d_x, t_x)$ stand for the estimated value from another individual lane, the weights are supposed to be direct

ratio to the rate of flows in a period:

$$\hat{q}_l(l_x, d_x, t_x) \sim \left({\bar{q}(l_x, d_x, t)}/{\bar{q}(l, d_x, t)}, t \in T \right) * (\bar{q}(l, d_x, t), t \in T_x) \qquad (4.18)$$

Then these values are weighted according to their spatial characteristics toward a final estimation. Weighting factors are introduced and all the coefficients for variables add up to 1. If lanes are not in a same phase group PL, the traffic streams on them are not in the same control, thus a zero is set. For the lanes in a same phase group PL, they are considered as equally importance:

$$w_{ph}(l) = \begin{cases} 0, & l \notin PL \\ {1}/{n_{pl}}, & l \in PL \end{cases} \qquad (4.19)$$

These aspects can be considered individually, too. The weighting factors of w_c and w_{tr} should be calibrated according to the specific situations. If there is no extra information from their side, then they are set as equal for each lane. The timing plan can still group the lanes by phase. In this case, the approach can be simplified as:

$$\hat{q}(l_x, d_x, t_x) \sim \sum_{l \epsilon PL} w_{ph}(l) * \hat{q}_l(l_x, d_x, t_x) \sim \overline{\hat{q}_l(l_x, d_x, t_x)}, l \in PL \qquad (4.20)$$

4.2.3. Approach 3: FCD - flow data fusion

The third approach starts from a different aspect, by linking external data source- FCD to the flow observations. It is a trial to find the possible contributions from external sources. Two areas near the junction are taken into consideration: the inbound area and the outbound area. In the inbound area, most of flows are interrupted flows influenced by the signal. In the outbound area, the flows are merged by other turning directions (straight flows from the opposite road segment, right turning and left turning flows from side segments). In each area, two relations are considered independently: the relation of FCD speed to loop flow and the relation of FCD count to loop flow.

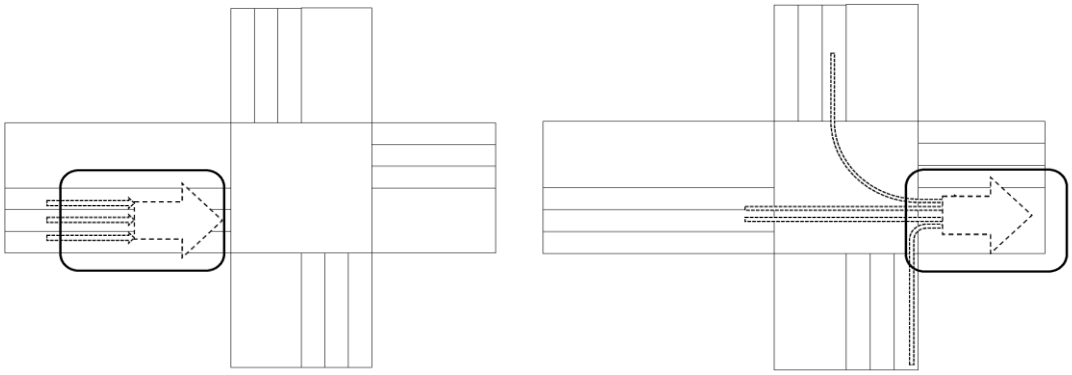


Figure 4-2 Two areas (in rectangles) to carry out data fusion inbound area (left) and outbound area (right). Arrows show the flow gathered from the detectors.

4.2.3.1. FCD speed-flow relation

The first way to conduct data fusion between FCD speeds and loop flows. An assumption is made that, a relation exists between the speeds and flows of a same traffic stream at an urban junction. If the speeds are obtained, and part of the flows are observed, then the rest flow values can be computed according to the relations found before. Although the speeds of the traffic streams are unknown, the FCD speeds act as representatives. The estimated flows in this way are aggregate flows.

Formulation

The steps are: Firstly, calculate FCD speeds to represent the speeds of traffic streams on a segment. Secondly, for a certain range, link available flows and the speeds, by fitting the two sources into one curve. Thirdly, apply the speed-flow curve to where the flows are missing, to get the estimated flow values. Finally, fill in the missing flows, update the curve g and go back to the second step. The formulations are:

$$u_{FCD}(L_x, d_x, T_x) \approx u(L_x, d_x, T_x) \tag{4.21}$$

$$u_{FCD}(L_x, d_x, T_x) \sim g(q(L_x, d_x, T_x)) \tag{4.22}$$

$$\hat{q}(L_x, d_x, T_x) \sim g^{-1}(u_{FCD}(L_x, d_x, T_x)) \tag{4.23}$$

Development

Although the speeds on corresponding segments are unknown, they are supposed to be deducted by other ways. In this work, available FCD helps to express the speeds on the segment.

A FCD record can be expressed as :

$$FCD(id, x, d, t, u, h) \tag{4.24}$$

id: The vehicle ID
x: Position,latitude and longitude
d: Day of recording
t: Time of day
u:Instantaneous speed of specific vehicle id at time t
h: Heading of specific vehicle idat time t

To conduct the mean speed u_n , each vehicle in a specific road segment L_x (which may contain multiple lanes) during period T_x is selected according to the headings. At the same time, they must match the road locations. Then, harmonic mean of speed $\overline{u_s}$ acts as the space mean speed.

$$u_n(id, L_x, d_x, T_x) \sim FCD(id, x, d_x, t, u, h), x = L_x, h \sim L_x \quad (4.25)$$

$$u_{FCD}(L_x, d_x, T_x) \sim \overline{u_s} = N / \sum_{n=1}^{N} \frac{1}{u_n} \quad (4.26)$$

4.2.3.2. FCD count-flow relation

Except for FCD speeds, FCD trajectory counts are used. FCD, as an independently measured data source, is part of the total flow. An assumption is made that, there are relations between FCD counts and loop flows. Linear curves are made to fit for these relations. If the counts are obtained, and part of the flow is observed, then the rest flow values can be computed according to the relations found before. Although the penetration rate of FCD changes according to different periods, the positive relations between FCD counts and the total flows can give some estimation. The estimated flows in this way are aggregate flows. And more advanced relations are expected in the future research.

Formulation

The steps of the algorithm are: Firstly, calculate FCD counts. Secondly, link available flows and FCD counts to a fitting curve. Thirdly, apply the counts-flows curve to where the flows are missing to get the estimated flow values. Finally, fill in the missing flows, update the curve g and go back to the second step. The formulations are:

$$k_{FCD}(L_x, d_x, T_x) \sim n(id, L_x, d_x, T_x) \quad (4.27)$$

$$k_{FCD}(L_x, d_x, T_x) \sim g(q(L_x, d_x, T_x)) \quad (4.28)$$

$$\hat{q}(L_x, d_x, T_x) \sim g^{-1}(k_{FCD}(L_x, d_x, T_x)) \quad (4.29)$$

Development

The counts of FCD are conducted by applying trajectories and sorting the same vehicle ID passing by an area in a certain period.

$$k_{FCD}(L_x, d_x, T_x) \sim n(id, L_x, d_x, T_x) \quad (4.30)$$

Other processes are similar to the algorithm using speed-flow relation described before.

4.2.4. Approach 4: Multiple linear regression (MLR)

The fourth approach makes the estimation by applying a multiple linear regression. In the first step, the potential contributions of observations to each others are assumed, the parameters showing these contributions are calibrated. In the second step, all these parameters are applied to available values to calculate the missing values.

Formulation

The formula is stated as follows, and the fitting function of the regression should be found. The sets of flows $q(l,d,t)$ are expressed as inputs x.

$$\hat{q}(l_x,d_x,t_x) \sim f_{regress}(q(l,d,t),\beta(l,d)), l \in L, d \in D, t \in T \quad (4.31)$$

$$f_{regress}(q(l,d,t),\beta(l,d)) = a(l,d) * x\big(q(l,d,t)\big) + b \quad (4.32)$$

n_{output}*: Number of output variables*
n_{input}*: Number of input variables*

Parameters a is a $n_{output} \times n_{input}$ matrix, and b is a vector with n_{output}. Parameter a and b are represented by β, so β is a vector of $n_{output} \times (n_{input}+1)$. A least squares multi-linear regression analysis is used to calibrate the parameters. It is done by minimizing the sum of the squared errors/residuals between the estimated values and the detected values:

$$\arg\beta \min E = \|\hat{q}(l,d,t) - q(l,d,t)\|^2 \quad (4.33)$$

Implementation

In the implementation of this approach, two important factors are there: the relevance of inputs and the analysis interval for parameters.

- *The relevance of inputs: Theoretically, all the data from one junction can be used. However, specific inputs should reply on the situation.*
- *The analysis interval: this factor determines how often the parameters are updated. Under an analysis interval of 8h, the weights of inputs keep unchanged during 8h.*

There is a dilemma when considering these two factors: On the one hand, more relevant inputs benefit to the accuracy of the estimation. For example, the flows from two nearby lanes have more commonalities. On the other hand, to capture the contributions from inputs, shorter analysis intervals are preferred. More frequent analysis leads to fewer inputs for each set of parameters of the regression, and the regression with more inputs requires more parameters. Thus, the analysis interval has to be long enough to ensure the accuracy of the parameters. Therefore, a trade-off must

be made between the inputs and the analysis interval.

Take a junction for example, which is with 6 lanes on each approaching stream (4 streams) observed for a week (7 days). If using all the observations at a junction as the inputs, the numbers of parameters are expressed as: A Constance + all the lanes multiplied with all the days in a week - 1(the missing flow itself), which is 1+4*6*7-1=168. More than 168 inputs are required, then the analysis interval has to be the whole day (5 min resolution, there are 288 sets of input). If choosing only the observations from one approaching stream, the numbers of independent variables have become 1+6*7-1=42. To ensure the reliability of parameters, the regression needs at least 42 equations. That is to say, the analysis interval should be at least 4 (48 sets of inputs) hours. The analysis interval can range from 4 hour to 24 hour.

4.3. Integration of the approaches

Each approach has specific concerns, advantages and limitations. The possible integrations are raised in this part to try to make improvements. Two integrated methods are: Combining approach 1 and 2 using iteration and advanced multiple linear regression.

4.3.1. Integrated method 1: Iteration

In this method, Historical pattern and lane spatial distribution are combined by involving the main dimensions from others. For example, when calculating the weights between days, not only the flows on this lane are considered, but also the lanes that are relevant. When calculating the weights between lanes, not only the flows on this day are considered, but all the relevant days. What is more, different from individual approaches, each output from one dimension can be the input of another dimension. The formulations are given by:

$$\hat{q}(l_x, d_x, t_x) = w_1 * \hat{q}_1(l_x, d_x, t_x) + w_2 * \hat{q}_2(l_x, d_x, t_x) \tag{4.34}$$

$$\hat{q}_1(l_x, d_x, t_x) \sim \{w_{DOW} * \bar{q}(l, d_{DOW_x}, t), w_{WD} * \bar{q}(l, d_{WD_x}, t), w_D * \bar{q}(l, d_{AD}, t), t \in T_x\}, l \in L$$

$$\hat{q}_2(l_x, d_x, t_x) \sim \sum_{l \epsilon L} w_c(l) * w_{tr}(l) * w_{pl}(l) \frac{1}{n_{pl}} \left(\frac{\bar{q}(l_x, d, t)}{\bar{q}(l, d, t)}, t \in T \right) * (\bar{q}(l, d, t), t \in T_x), d \in D$$

Where w_1 and w_2 are calibrated during the iteration. The Iteration process can be expressed by the flow chart and explained by the following steps.

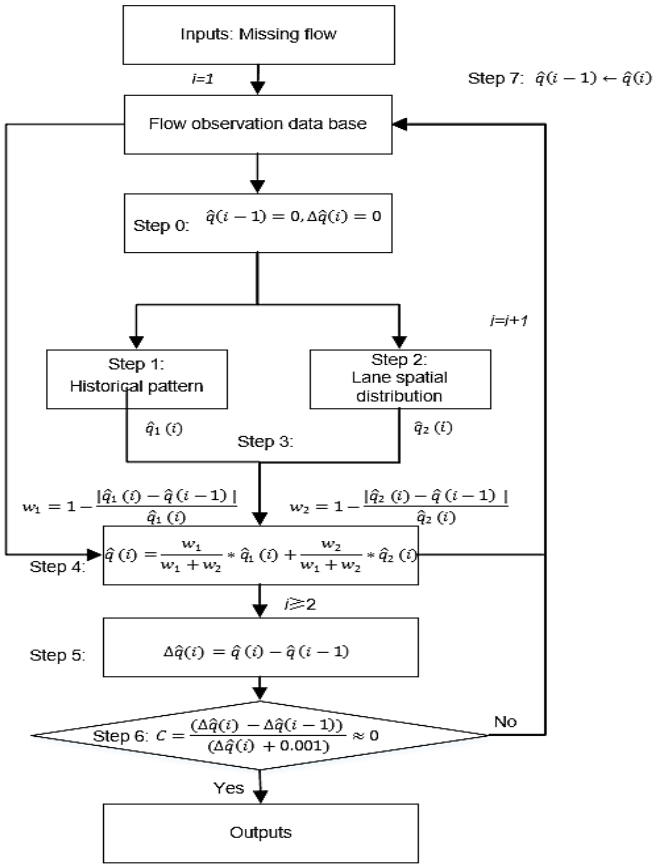


Figure 4-3 Schematic flow chart of the Iteration

Step 0: For initial condition, the estimated value is $\hat{q}(0) = 0$. Set the deviation of estimated values between the last step and this step as $\Delta\hat{q}(0) = 0$;

Step 1: Get estimation value $\hat{q}_1(i)$ by historical pattern approach;

Step 2: Get estimation value $\hat{q}_2(i)$ by lane spatial distribution approach;

Step 3: Calculate the reliability weighting factor for two independent estimated results:

$$w_1 = 1 - \frac{|\hat{q}_1(i) - \hat{q}(i-1)|}{\hat{q}_1(i)}, w_2 = 1 - \frac{|\hat{q}_2(i) - \hat{q}(i-1)|}{\hat{q}_2(i)} \quad (4.35)$$

Step 4: Calculate estimated value in i term, based on weighted average of each component:

$$\hat{q}(i) = \frac{w_1}{w_1+w_2} * \hat{q}_1(i) + \frac{w_2}{w_1+w_2} * \hat{q}_2(i) \quad (4.36)$$

Step 5: Calculate the deviation of i term from last term i-1:

$$\Delta\hat{q}(i) = \hat{q}(i) - \hat{q}(i-1) \quad (4.37)$$

Step 6: Calculate the convergence, if converge, stop; if not, go to step 8.

$$C = (\Delta\hat{q}(i) - \Delta\hat{q}(i-1)) / (\Delta\hat{q}(i) + 0.001) \approx 0 \quad (4.38)$$

Step 7: Update the missing flow $\hat{q}(i-1) \leftarrow \hat{q}(i)$ in the observations and go back to step 1.

4.3.2. Integrated method 2: Advanced multiple linear regression

This method is based on approach 4, using regression tools. The information of relevance comes from the considerations in approach 1 and 2. This means, historical pattern and lane spatial distribution in a timing plan have given the regression model the guidance to selecting its inputs. The formulation is:

$$\hat{q}(l_x, d_x, t_x) \sim f_{regress}(q(l,d,t), \beta) \quad (4.39)$$

$$l \in PL, TL, CL, d \in DOW_x / WD_x AD, t \in T$$

The way to find the function is still the same as approach 4.

$$f_{regress}(q(l,d,t), \beta) = a(l,d) * x\big(q(l,d,t)\big) + b \quad (4.40)$$

$$\arg\beta \min E = \|\hat{q}(l,d,t) - q(l,d,t)\|^2$$

4.4. Conclusion for the chapter

In this chapter, approaches to estimate missing flow are raised from two aspects: direct observation and traffic flow theories using data fusion. Based these two aspects, four individual approaches and two integrated methods are proposed. The first approach, historical pattern, uses the pattern over the time dimension to make an estimation. Lane spatial distribution, as the second approach, uses the similarity of flows over the space dimension. FCD and loop flow are combined in approach 3. Both speeds-flows and counts-flows relations are used to make the estimation. MLR is applied in the approach 4. Except for the formulations, the way to consider inputs and analysis interval is also introduced. For the first integrated method, an iteration is introduced to combine the first and the second approach. In the second integrated methods, the MLR is improved by considering its inputs according to information from the first and the second approach. The performances of these approaches and methods are demonstrated in Chapter 6.

5. Data processing and implementation of methods

The data processing is a significant part in the thesis work. The outputs are useful in data analysis and the experiments in case studies. Therefore, Chapter 5 introduces these processes, which including: the coding of junctions, the indexing of flows, the timing plan processing, and the processing of FCD trajectories. Some technical improvements are introduced, such as a new coordinate system to form FCD trajectories.

5.1. Loop flow and timing plan data processing

SCATS saves its traffic flows and signal timing plan data in files every day. The SCATS traffic reporter gives the raw data. The two raw data files are SCATS_VS_Flow and SCATS_SM_Timing. The former one provides the flows over every 5 min, and the later one provides the timing plan and relative information each cycle. SCATS data are available from 2013.4.15 to 2013.4.28. The file on 2013.4.22 Monday has broken. To complete the two-week make experiment range, data from the next Monday (on 2013.4.29) are used. To extract all the values and to form a dataset, there calls for geography information, such as the positions of lanes, the turnings. All these have to be done by the junction coding.

Coding of junctions

The first step is to check current junction information and revise the errors according to the Map from SCATS and the Google map. In previous year table, some of the information of junctions is wrong or incomplete. For example, for some junctions, the numbering of the junctions doesn't match with the numbering in SCATS. Besides, for some junctions, there are no numbering. With the completion of the junctions' information, it has been more convenient to refer to traffic flow observations. Here some corrections of the junction's information are shown in yellow.

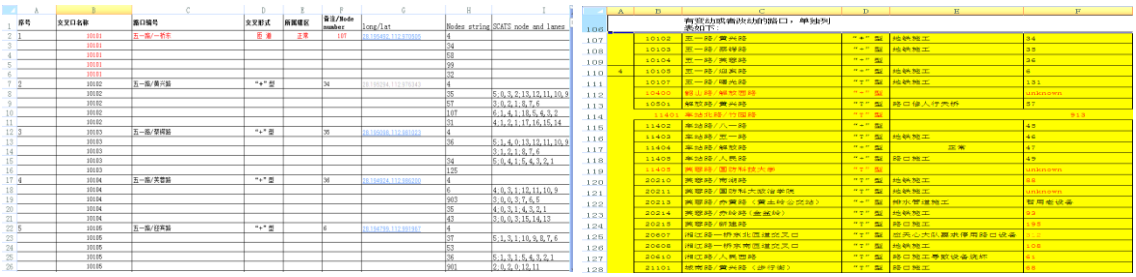


Figure 5-1 Example of junctions (left) and check of errors in junction information (right)

The next step is to extract flow values and to create a database. By sorting and classification, data are organized into groups, labeled with the day and lane coordinates, matched with group information and timing plan. After the processes, each flow observation is noted with a junction number, a date, and a lane number.

Timing plan process

The raw data of signal timing plan contain information about the groups, the duration of phases and the starting time of cycles. The groups of lanes have been introduced in last part. For the duration of phase, they are extracted from raw dataset and matched with the flow recording interval (5 minutes). An example is presented here from 7:00 to 7:30 on a day. In this diagram, a rectangle shows a lane on the vertical axis, while the horizontal axis shows the time line. Each red square represents one minute. Yellow

lines show the starting time of cycles, and yellow numbers show the durations (in second). Cyan numbers in cyan rectangles show the greens (in second).

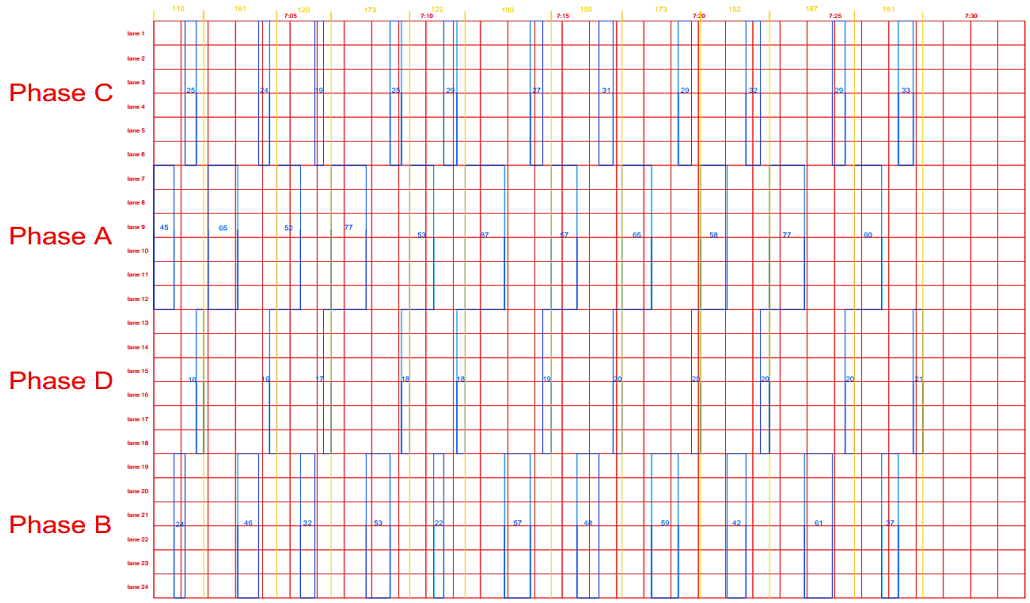


Figure 5-2 Sample of timing expression; flow values are involved within 30 minutes

There is only a starting time and a duration time for each phase. The yellow lights and all-red lights have to be assumed by shifting the rectangles for a synchronization during a time range. Afterwards, the timing data for each lane at each day are saved.

The greens are shown on an aggregated 30-minutes level. It shows that the greens in each phase keep unchanged during the daytime, but fluctuate during the night. The reasons are supposed: (1) SCTAS did not apply an adaptive control according to the flows during the daytime. (2) During the daytime, the traffic flows on four approaching directions increase or decrease on a same scale, thus their relative ratios keep unchanged, which leads to an unchanged share of greens .

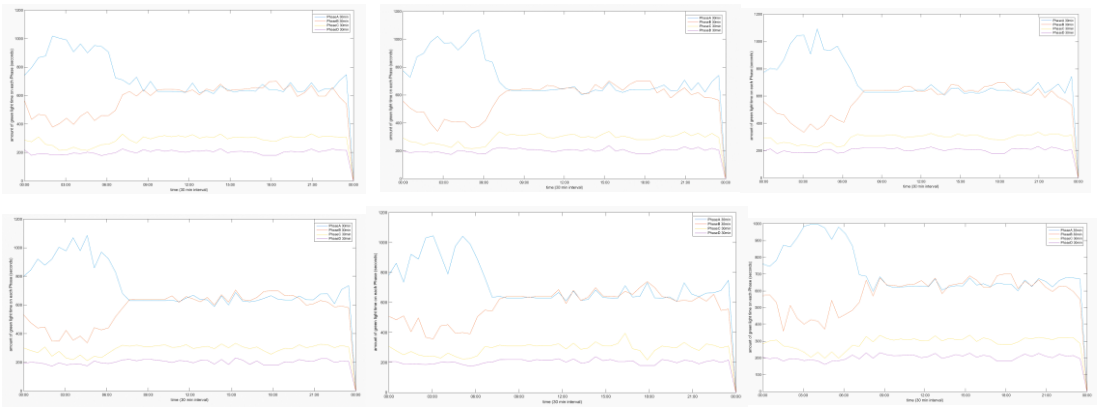


Figure 5-3 the green light time distribution, each phase on 15th-20th April 2013(30 minutes interval))

5.2. FCD processing

FCD, as another data source used in the thesis work, not only plays a role in verifying the relations in traffic streams, but also gives a contribution to the consistency check. To get information, FCD trajectories should be formed from the raw data. Two kinds of trajectories are there. The first type is a trajectory from the overhead view; it uses two location coordinates: longitude and latitude. These trajectories give counts of vehicles through a road segment. The second type uses distance and time as coordinates. These trajectories can be matched with a timing plan in a short period, and can be used for the calculation of the average speed.

First type trajectory

The first type of FCD trajectory provides an overhead view of each taxi. The author uses these trajectory to check the FCD coverage and to calculate the vehicle counts in a certain location. If putting GPS location and SCATS junction location onto the Google map, it is obvious that they are not matched. Thus, a map-matching process should be conducted. Compared to FCD on freeways, the FCD near junctions holds several difficulties: These vehicles are no longer heading only to two directions as on the freeway, but are in multiple directions. For a regular junction with four approaching directions, there can exist as many as 24 FCD trajectories with different headings. For irregular junctions, there can be much more. The ways to select the headings are introduced in following parts.

Second type trajectory

The second type uses distance and time as coordinates. The majority of vehicle trajectories have shown more reasonable after been implemented PLSB (Piecewise linear speed based method). However, some vehicles show strange behaviors such as U-turns. Current methods cannot filter out those U-turns completely. Luckily, these special cases have not influenced the analysis of traffic state at junction using FCD trajectories nor the collection of vehicles counts to a large degree. Near a junction, a shockwave can be observed, which helps to check the consistency of trajectories and the signal plan. The queues form back forwards when a red signal is on. When the signal turns green, the flow reaches its capacity on this road; vehicles pass the junction and the queues start to dissolve. Here is an overview of all taxi trajectories:

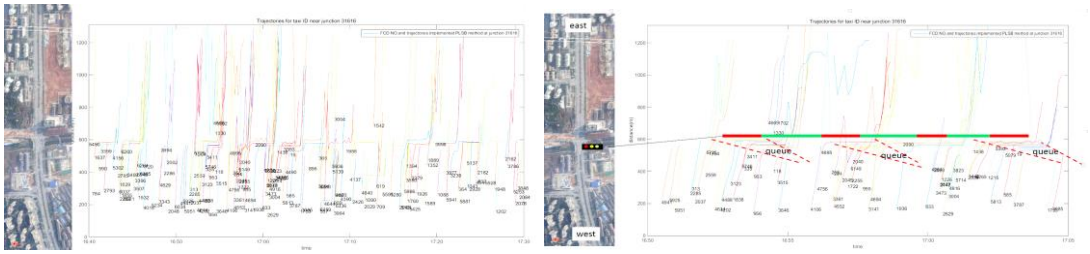


Figure 5-4 Example of FCD trajectories near a junction 31616 from west to east from 16:50 to 17:05 compared with control plan.

Short of FCD vehicles

To form both trajectories, the sort of vehicle is the main task. All records from raw data are classified by their Taxi ID, followed by the time. Since specific vehicles and trips have been determined, the headings are concerned, with their locations. These vehicles are then matched with approaching directions at a junction. A sheet of sorted data with their trip information is shown as follows (the red circle shows the trips for a specific vehicle ID).

ID	time	speed	lon	lat	X	Y	Z	heading	triprecord
1	[7.3060e...	[0;4.1000;...	[113.017...	[28.1604;...	[-2.1987e...	[5.1755e...	[3.0000e...	[310;0;170;...	[1,3]
2	[7.3060e...	[0;43.200...	[113.017...	[28.1605;...	[-2.1987e...	[5.1755e...	[3.0000e...	[300;270;2...	[1,2,6]
5	*23x1 dou...*	*23x1 dou...*	*23x1 dou...*	*23x1 dou...*	*23x1 dou...*	*23x1 dou...*	*23x1 dou...*	*23x1 double*	1
6	7.3060e+...	25.7000	113.0178	28.1603	-2.1988e...	5.1755e+...	3.0000e+...	260	1
8	*15x1 dou...*	*15x1 dou...*	*15x1 dou...*	*15x1 dou...*	*15x1 dou...*	*15x1 dou...*	*15x1 dou...*	*15x1 double*	[1,2]
9	*14x1 dou...*	*14x1 dou...*	*14x1 dou...*	*14x1 dou...*	*14x1 dou...*	*14x1 dou...*	*14x1 dou...*	*14x1 double*	[1,4,12]
11	[7.3060e...	[69.5000;...	[113.020...	[28.1606;...	[-2.1990e...	[5.1753e...	[3.0000e...	[80;90;80;1...	[1,4]
13	[7.3060e...	[0.2000;2...	[113.023...	[28.1604;...	[-2.1993e...	[5.1752e...	[3.0000e...	[350;340;2...	1
14	[7.3060e...	[17.6000;0]	[113.023...	[28.1615;...	[-2.1992e...	[5.1752e...	[3.0001e...	[170;180]	1
17	[7.3060e...	[25;48.50...	[113.023...	[28.1611;...	[-2.1992e...	[5.1752e...	[3.0001e...	[200;260;2...	[1,3,9]
18	7.3060e+...	25.4000	113.0234	28.1611	-2.1992e...	5.1752e+...	3.0001e+...	170	1
19	*16x1 dou...*	*16x1 dou...*	*16x1 dou...*	*16x1 dou...*	*16x1 dou...*	*16x1 dou...*	*16x1 dou...*	*16x1 double*	[1,4,10,12]
20	*11x1 dou...*	*11x1 dou...*	*11x1 dou...*	*11x1 dou...*	*11x1 dou...*	*11x1 dou...*	*11x1 dou...*	*11x1 double*	[1,3,11]
22	[7.3060e...	[58.5000;...	[113.019...	[28.1605;...	[-2.1989e...	[5.1754e...	[3.0000e...	[70;80;90;0...	1
25	*15x1 dou...*	*15x1 dou...*	*15x1 dou...*	*15x1 dou...*	*15x1 dou...*	*15x1 dou...*	*15x1 dou...*	*15x1 double*	[1,4,6,7,8,12,...
26	7.3060e+...	56.3000	113.0202	28.1605	-2.1990e...	5.1754e+...	3.0000e+...	80	1
28	*12x1 dou...*	*12x1 dou...*	*12x1 dou...*	*12x1 dou...*	*12x1 dou...*	*12x1 dou...*	*12x1 dou...*	*12x1 double*	[1,9]
30	[7.3060e...	[38;74.60...	[113.024...	[28.1606;...	[-2.1993e...	[5.1752e...	[3.0000e...	[60;80;270;...	[1,3,4]
31	[7.3060e...	[0.2000;0;...	[113.023...	[28.1605;...	[-2.1993e...	[5.1752e...	[3.0000e...	[350;350;3...	1

Figure 5-5 Example of FCD data-taxi ID and trips classification

For each user, there could be traces in an area during different periods. A threshold is used to sort the trips:

$$threshlod \sim max\, t_{interval} \tag{5.1}$$

The threshold is set as 5min: $max\, t_{interval}$=5min. It means that if the interval between two consequence records has exceeded 5min, these two records are thought to be independent trips.

New coordinate system

Sorted vehicles should be matched to the roads on the map. Due to the lack of link-match sources, the link-match and the turning determination have to be done manually. Since the targets are the roads near junctions, not only two headings are there when considering the movements of vehicles. For example, vehicles may make turnings at junctions. Moreover, the shapes of the junctions may also influence the trajectories. Thus, a new coordinate system is invented by the author to consider these issues.

To form the new coordinate system, the first step is to check the headings of FCD. After tests and checks, the original coordinate system used in the FCD raw data is confirmed (as shown in Figure 5-6 (left) figure). Based on the original one, a new coordinate system is developed. In the new coordinate system, each record point, with their actual longitude and latitude is projected to the uniformed axis at this junction. Four boundary points are decided automatically by the median value of the vehicles near this boundary, thus the shape of the coordinate system can adjust automatically to the shape of the targeted junction.

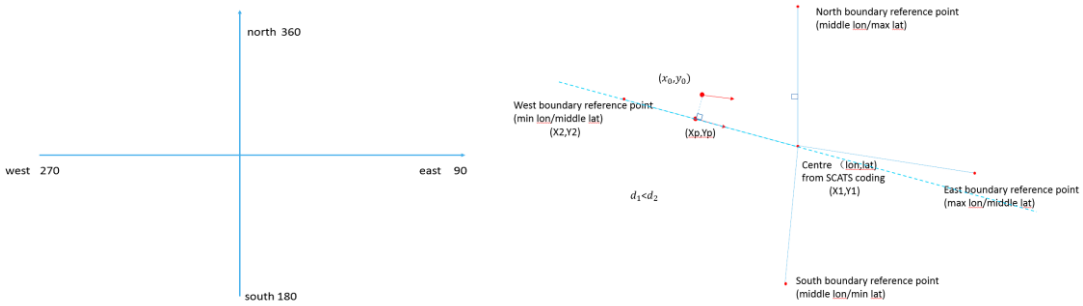


Figure 5-6 The original coordinate system in FCD (left), newly designed coordinate system (right)

In the new coordinate system, reference lines are set to link four boundary points and the junction center (see the blue solid lines in the Figure 5-6 (right)). For the points on the west branch of the junction, the reference line is decided by the west boundary reference point and the center reference point, represented as $Ax + By + C = 0$. The aim of the reference line is to make projections for all origin recorded points. So that it will help to calculate and match the points to a certain link. Set the position of a vehicle as (x_0, y_0). The parameters A, B and C are decided by the position of junction center and one of the four boundary reference points. Thus, the projections of the origin recorded points will be on this line, and they are given by:

$$x_p = (B * B * x_0 - A * B * y_0 - A * C)/(A * A + B * B) \quad (5.2)$$
$$y_p = (-A * B * x_0 + A * A * y_0 - B * C)/(A * A + B * B)$$
$$A = \frac{y_2 - x_2}{y_1 - x_1}, B = -1, C = y_1 - x_1 * (\frac{y_2 - x_2}{y_1 - x_1})$$

Where x_p and y_p are the new longitude and latitude to calculate the distance between two records.

The blue dotted line in the figures represents the dynamic reference line for a vehicle at a different position. The distance from the recorded point to the link is given by:

$$d = (A * x_0 + B * y_0 + C)/\sqrt{(A * A + B * B)} \quad (5.3)$$

If the vehicle goes straight, d_1 will be smaller than d_2 all the time; at the same time, the heading will not change much. However, if the vehicle turn right at the junction, at first d_1 will be smaller than d_2, then it becomes larger; at the same time, the

heading will change obviously. The table Table 5-1 shows this determination.

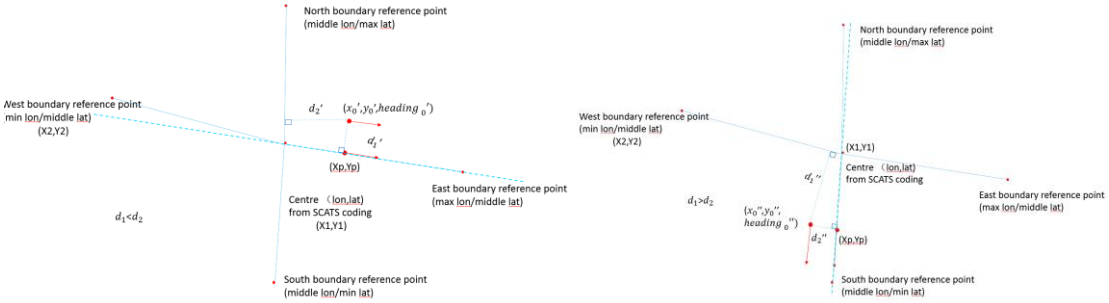


Figure 5-7 the FCD coordinate designed for data processing of FCD heading.

Table 5-1 Determination of turning in new coordinate

Heading	Distance to reference line 1/ line 2		Determination
West-west	$d_1 < d_2$	$d_2 < d_1$	From west to east
West -south	$d_1 < d_2$	$d_1 > d_2$	From west to south

This method is proved to be able to deal with the majority of the trajectories. However, it still needs to be improved in the future.

Distance calculation using PLSB

To get the distance between two records of a vehicle, the original longitude and latitude information should be transferred. The Euclidian distance between two points is widely used, since it is easy to calculate. To calculate the Euclidian distance, a reference system is needed. WGS84 (World Geodetic System 1984) is a reference system, which is often used as a base to make maps and calculate distances. It determines the coordinates of every point on the earth using latitude, longitude and height. A schematic diagram is shown in Figure 5-8.

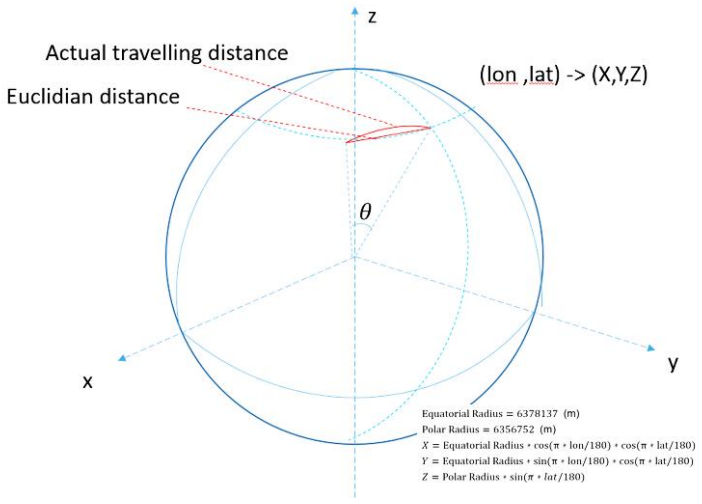


Figure 5-8 WGS84 used in FCD data processing

The Euclidean distance between two points:

$$D = \sqrt{(X_1 - X_2)^2 + (Y_1 - Y_2)^2 + (Z_1 - Z_2)^2} \quad (5.4)$$

As long as the angle θ between the start point and the end point is small enough, the Euclidian distance will be close enough to the actual geographical distance. For example, for a taxi with a maximum speed of 40 km/h or 60km/h, in an interval of 30 seconds, the maximum distance of two recorded points will not exceed 500 m.

Having decided the distances between points, the next task is to arrange these distances. Since the original data could show vague trajectories of a vehicle; for example, some shapes of the trajectories could be very 'sharp'. To smooth the trajectories, the PLSB method is used. It stands for piecewise linear speed based method for travel time prediction. In this method, trajectories are constructed based on the assumption of piecewise linear (and continuous at section boundaries) vehicle speeds. The following figure shows how the PLSB is implemented.

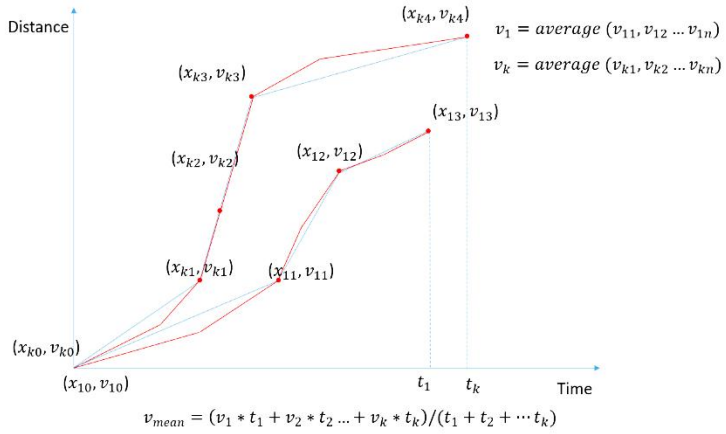


Figure 5-9 Example of forming of PLSB method to a single vehicle

5.3. Conclusion for the chapter

This chapter demonstrates the complexity of data and how to process raw data. Firstly, a digital coding of junction information is made using SCATS documents, together with maps. Secondly, timing plan is utilized. Thirdly, attributes from FCD are used to generate two kinds of trajectories, providing speeds and counts. In the data processing FCD, a new coordinate system is invented. Errors and inconsistencies are concerned, too. The data filtered are useful for next step usage.

6. Case studies and evaluation

This chapter presents the experiment results and corresponding evaluations. Two junctions are used, and various methods are applied at the first junction for testing and calibration purposes, and two integrated methods implemented with the second junction for purpose of validation.

Section 6.1 describes the general setup for each of the cases being considered, including the key influencing factors in each of the scenarios, the indicators used in evaluation, and the choice of junctions.

In section 6.2, the first approach, which is based on an historical pattern, contains two sub-approaches. In sections 6.3 to 6.5, approaches 2, 3, and 4 are implemented and evaluated. In sections 6.6 and 6.7, integrated methods using iteration and an advanced regression method are conducted and evaluated. Section 6.8 presents the validation results by using another junction to show the possibility of using methods extensively; that is to say, with robustness. Section 6.9, the conclusion, includes a summary of all of the methods.

6.1. Setup for case studies

Figure 6-1 presents a flow chart of the processes described in this chapter. In some cases, (identified here as approaches 1.1 and 2), initial calculations following certain algorithms belonging to this approach are conducted to indicate the trends in certain performances. The updated algorithms are then used.

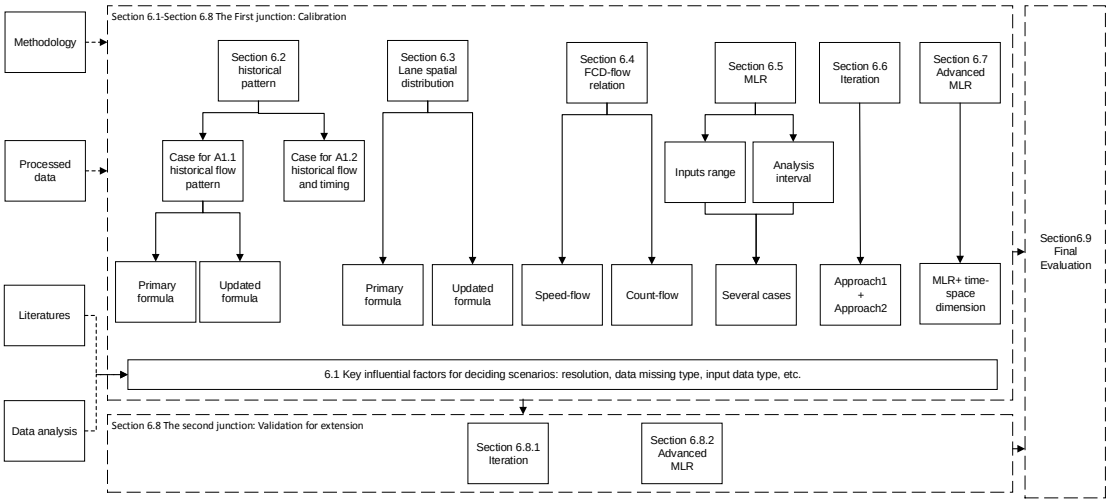


Figure 6-1 A flow chart of the discussion in the case study and evaluation chapter

The case studies of the methods are carried out in multiple scenarios, and key influencing factors define the setup of these scenarios. Thus, this section will introduce the key factors that were considered in determining the methods to be followed and the setup of cases and experiments.

Key influential factors

Based on the current imputation methods and the results from the data analysis, certain factors are considered as influencing the performance of the methods. These are:

- Missing data: type — long-term or short-term.
- Input data type: original data (raw data taken directly from detectors) or processed data (smoothed data)
- Resolution: 5, 15, or 30 min.
- Analysis interval (frequency of capturing the rules and calibrating the parameters for approach 4 at: 4h, 8h, 12h, and 24h.

The moving average, a widely used tool, is applied for smoothing the original data. A simple moving average (SMA) is the unweight mean of the previous n values (6.1). When considering the data before and after observations, the formula is (6.2). Setting n as 1, a simple way to smooth the raw data is (6.3).

$$q_{SMA}(t) = \frac{q(t)+q(t-1)+q(t-2)+q(t-(n-1))}{n} \quad (6.1)$$

$$q_{SMA}(t) = \frac{q(t)+[q(t-1)+\cdots+q(t-(n-1))]+[q(t+1)+\cdots+q(t+(n-1))]}{n*2+1} \quad (6.2)$$

$$q_{SMA}(t) = \frac{q(t)+q(t-1)+q(t+1)}{3} \quad (6.3)$$

All of the analyses in the following approaches are based on taking as inputs the original data and the processed data.

Setup of cases and experiments

- ***Error indicators***

The evaluation indicators are MAPE (mean absolute percentage error) and RMSE (root-mean-square deviation). These are defined in the following formulas. In the equation, estimated values are $\hat{y}_t$, while the actual value is y_t.

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\hat{y}_t - \frac{y_t}{y_t}\right| \quad (6.4)$$

$$RMSE = \sqrt{\sum_{t=1}^{n}(\hat{y}_t - y_t)^2 / n} \quad (6.5)$$

- ***Junctions in use***

Some junctions with full flow observations are selected from the system. Selected flows are "removed" from their places and reserved as the actual detected data for comparison. There will "missing" flows at a particular location (lane l_x) on a particular day (day d_x) during a particular time period (time of day $t_x \in T_x$).

The first junction is that of Road Wanjial and Road Laodong in Changsha, China, and is marked as ID 31616 at Changsha 2nd term, in the SCATS system. The second junction is that of Road Wangjiali and Road Qutang, with ID 31617. Their layouts are shown in Figure 6-2.

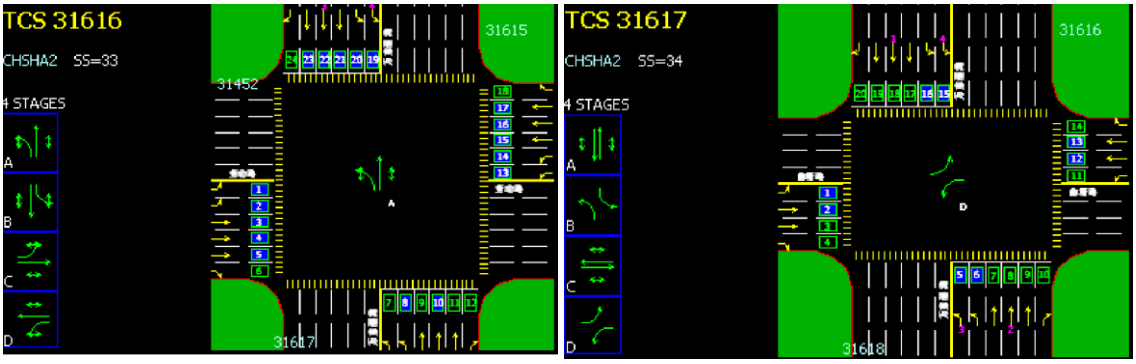


Figure 6-2 Layout of case junctions in SCATS, junction 31616 (left) for calibration and evaluation, and 31617 (right) for validation.

Junction 31616 has four approaching directions. At each approach there are, from left to right: two left-turning lanes, three straight lanes, and one right-turning lane, respectively. Junction 31617 has four approaching directions, too. In the east and west directions, there are four lanes each. In the south and north directions, there are six lanes. Since FCD covers these areas, sufficient values can be obtained as a reference.

- ***Data scope***

SCATS data are available from 4/15/13 to 4/28/13. The file from Monday, 4/22/13 broke. To complete the two-week experiment range, data from the next Monday (4/29/13) are used.

- ***Cases setup***

There are many ways in which permutations of the influential factors can be performed. To avoid too many permutation cases, one or two factors are identified for convenience. For example, it is recommended to keep all inputs at a 5 min resolution and switch the missing data type from long-term to short term, or else to switch the input data type from original data to smoothed data.

- ***Validation setup***

To validate the approaches or methods, some rough standards are made. In the way of validating, the methods are implemented in other similar lanes in another comparable junction to see the errors. As for the acceptable range of the difference of performance, for the convenient, it is defined that: the general difference of error are within 20% of the total error, the methods are seen as validated.

6.2. Cases for approach 1 Historical pattern

This approach involves two sub-approaches: Approach 1.1 uses only historical flow, and the earlier algorithms and upgraded ones are both demonstrated. Approach 1.2 uses the ratio of flow to green as the relevant variable to be identified in the history.

6.2.1. Historical flow pattern

In this section, the simple and primary formulas from the early experiment are first presented to show general performances. An updated version is then carried out with a specific lane to provide detailed results.

The primary algorithms: general performances

Primary algorithms use an average of flows from WDs (days in the same week) and DOWs (days same as day-of-week).

$$\hat{q}(l_x, d_x, t_x) \sim \bar{q}(l_x, d, t_x), d \in WD_x \tag{6.6}$$

$$\hat{q}(l_x, d_x, t_x) \sim \bar{q}(l_x, d, t_x), d \in DOW_x \tag{6.7}$$

Only the error indicators are presented, to show changes in performance over a number of days in three lanes. Here the red line shows WDs, while the yellow line shows DOWs.

- ***MAPE***

It can be seen that the percentage errors fluctuate over different streams in various directions. However, their trend on each day is similar: on Saturday and Sunday, all of them show higher errors, while on Thursday, they all reach an optimum performance. This yields the implication that the use of historical flow patterns relies on a stable rate of flow over a series of days.

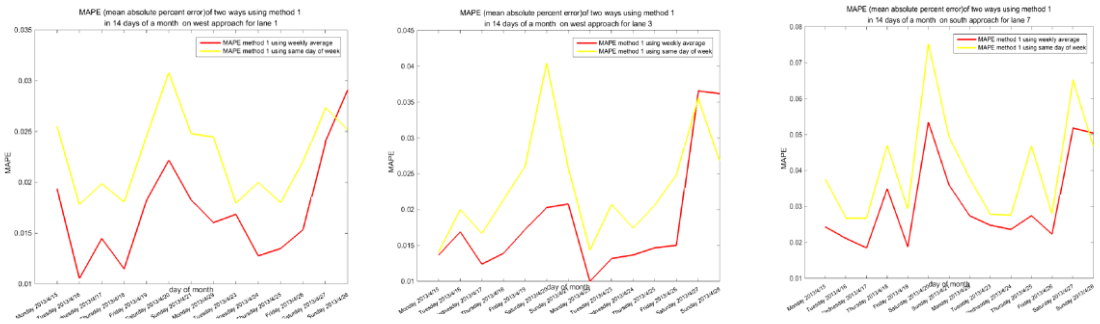


Figure 6-3. MAPE on the stream level, using an historical flow pattern over a two-week period, based on missing flow data on (1) West stream lane 1 (2), West stream lane 3 (3), and South stream lane 7.

From the results of MAPE, It should be noted that the MAPE in this section is calculated by comparing aggraded flow to the whole stream; thus the results may seem optimistic.

- ***RMSE***

RMSE yields an estimation from an absolute deviation perspective. It fluctuates in the same manner as MAPE. As the South stream holds a larger traffic volume, it also provides the largest error. It can be concluded that the absolute error of using an historical flow pattern increases as the increase in the amount on the stream where the missing flow lane is located.

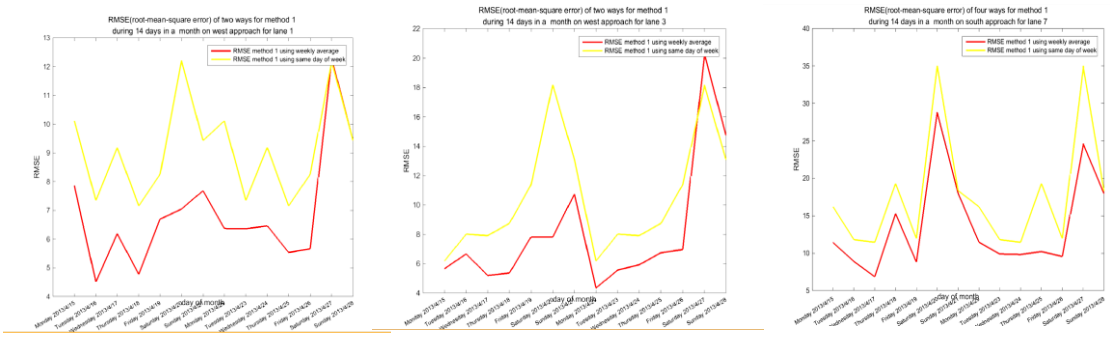


Figure 6-4. RMSE at the stream level over two weeks, based on missing flow date for (1) West stream lane 1, (2) West stream lane 3, and (3) South stream lane 7.

Considering both indicators, using the average of flows from WDs shows lower error indicators than when using flows from DOWs. However, this conclusion should be considered tentative, since the former has a larger amount of data input than the later, due to the limitation of data scope in the experiment.

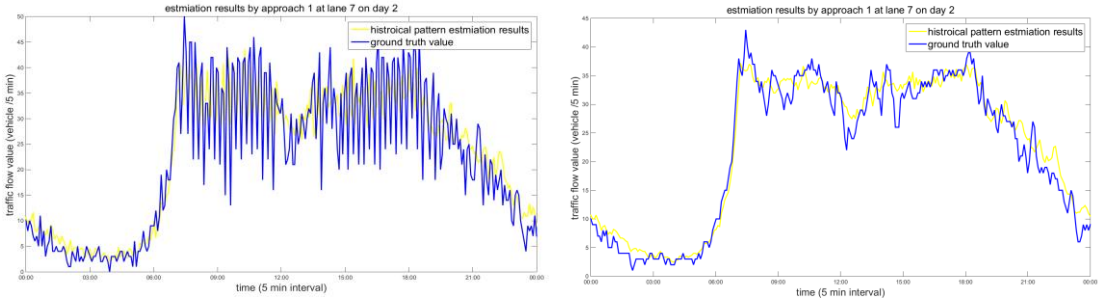
The updated algorithm

Having obtained general performances using primary formulas, the algorithms are updated by combining the influence from several relevant groups of days. The fractional part on the left side shows the DOW ratio, which refers to the general traffic demand of DOW over DOWs. The right side shows the average flow level in these cycles within a week.

$$q(l_x, d_x, t_x) \sim \frac{\bar{q}(l_x, d, t_x), d \in DOW_x}{\bar{q}(l_x, d, t_x), d \in AD} * (\bar{q}(l_x, d, t), d \in WD_x, t \in T_x) \tag{6.8}$$

The general performances of the whole junction are presented in Appendix 1. The tables in Appendix 1 reveal that the relative errors are stable over different cases (missing flow in different lanes on different days). In some cases, the process of smoothing largely improves the performances. Compared to left-turning lanes, estimations using throughput lanes enjoy higher accuracy. Also, compared to lower demand-approaching streams (East-West), higher demand ones (North-South) give better performances.

Pick up one case for evaluation, the case involves the South stream in week 2, day 2 April 23, 2013, in lane 7. (The yellow lines show the estimated values while the blue lines show the actual ones).



Error indicator (5min interval)	original data	processed data
MAPE	32.59%	17.00%
RMSE	6.46	2.62

Figure 6-5. Estimation using approach 1.1 on lane 7 junction 31616, April 23, 2013, for original data (left) and processed data (right).

This approach successfully captures the general trend. However, the peak captured from other days seems to "shift" a little bit when looking at the original data. For the processed data case (smoothed flow volume), the error indicators are much lower.

The findings in the results of approach 1.1 (A1.1) show that historical flow patterns fluctuate wildly among different cases. The results from primary algorithms are around 10% and 40%. The upgraded algorithms show a performance of around 30% using original data and around 15% with processed data. The approach relies on a stable flow rate for a series of days. It turns out that even for days with stable flows, results using this approach may "shift" a little bit in their time dimension from the actual values. The absolute errors increase with increases in the flow amounts. Also, due to the limitations of data scope in the experiments, there is no evidence as to which sets of days, DWs or DOWs, are more reliable.

6.2.2. Historical timing + flow pattern

In the implementation of this approach, to avoid fluctuations in a control cycle (which have a very short period of between 2 and 3 min.), the relationship between green and flow observations is established for a longer period (5 to 60 min.). This experiment is conducted on an aggregate level of 30 minute intervals. The formula used in the experiment is:

$$Q(l_x, d_x, t_x) \sim s(l_x, d_x, T_x) * \frac{\bar{r}(l_x, d, t_x), d \in DOW_x}{\bar{r}(l_x, d, t_x), d \in D} * (\bar{r}(l_x, d, t_x), d \in WD_x) \quad (6.9)$$

The green light/flow ratio

The green/flow ratios in Phase C (where lane 1 is located) over the course of a week are put together: it seems that the ratios during the night are quite close to each other.

However, the ratios for the daytime fluctuate widely.

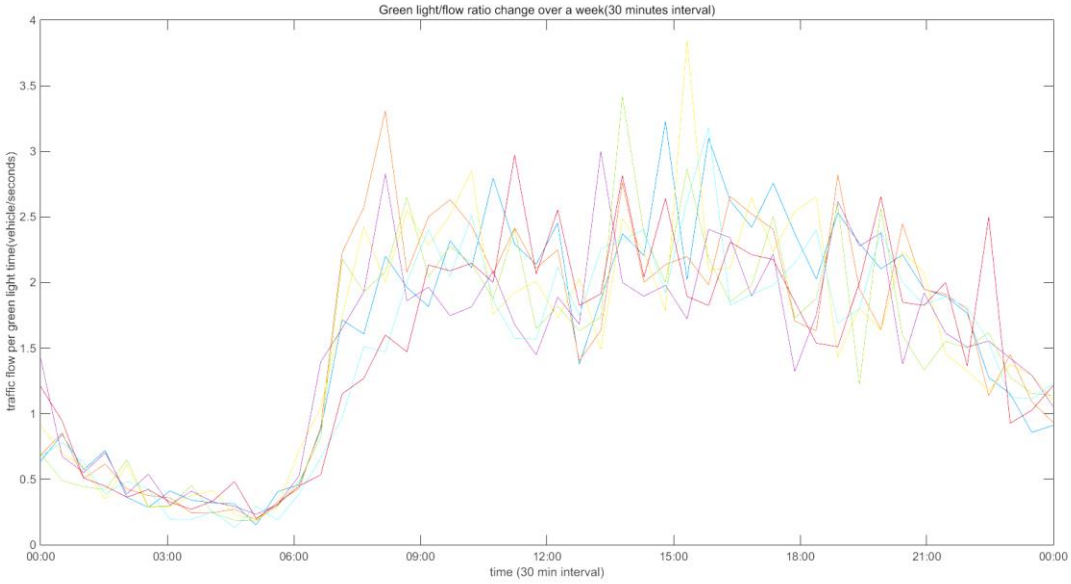
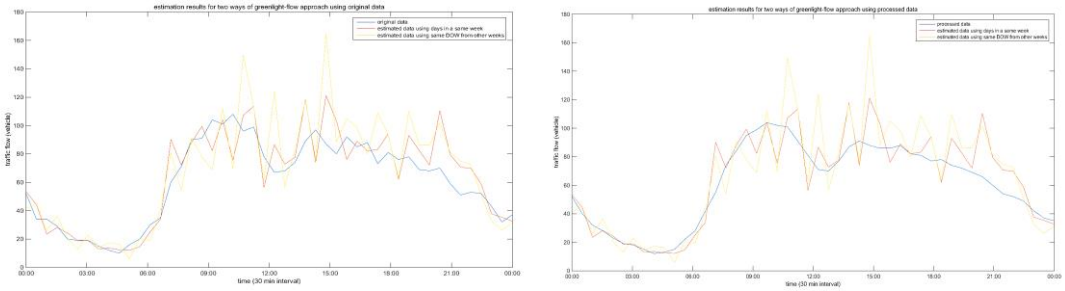


Figure 6-6. The green/flow ratio over one week (April 15-21, 2013) at 30-minute intervals.

Estimated flows

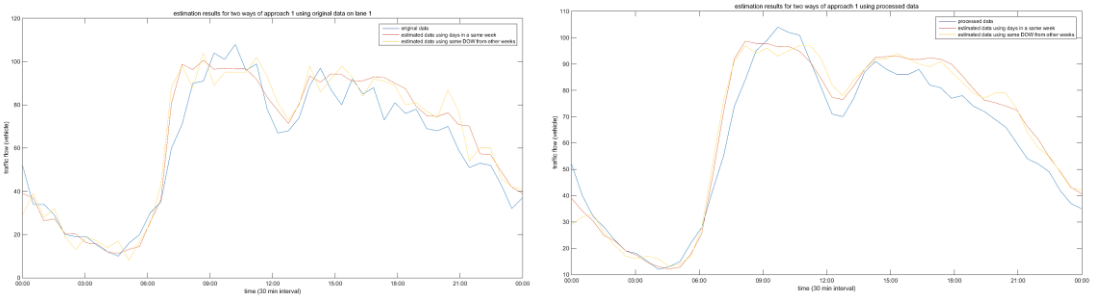
The green/flow ratios are used to calculate the missing flows. If looking back to the respective shapes of signal timing plans and traffic flows, it can be seen that due to the unchanged green, the final results are derived mainly from the historical traffic flows.



Error indicator(30min interval)	original data	processed data
MAPE	18.33%	15.25%
RMSE	16.54	14.96

Figure 6-7. Estimated results using the green light time/flow ratio approach in lane 1 on April 15, 2013: original (left) and processed data (right) with 30-minute interval.

In this case, that approach is not suitable. Thus, the greens are removed, and the approach falls back to 1.1, using only the historical flow. The results are:



Error indicator(30min interval)	original data	processed data
MAPE(the whole day)	12.07%	10.32%
RMSE(the whole day)	8.97	7.40

Figure 6-8 estimated results using degenerated approach 1.2 on lane 1 on day 15th April 2013-original data (left) and processed data (right) with 30 minute interval

According to the experimental results, the greens in phase C remained almost unchanged during the entire day. For this reason, the ratios of green to flow follow the same shape as the flows. The final results are almost entirely contributed by the flow values. Without the information on historical greens, historical traffic flows can produce even better results. Therefore, the Approach 1.2 (A1.2) of historical timing and flow pattern does not show its utility in this thesis, due to the unchanged green in the daytime. It can be concluded that this method is not suitable when a system does not adapt its controls according to the flows. This approach thus automatically falls back to approach 1.1, using only the historical flows.

6.3. Cases for approach 2 Lane spatial distribution

As with approach 1, first, early results using the primary algorithms are demonstrated. Then a more specific and updated version is implemented in order to provide detailed results.

The primary algorithms: general performances

First, the average PL (lanes in the same control group) is used. Secondly, the average TL (flow from lanes turning in the same direction):

$$\hat{q}(l_x, d_x, t_x) \sim \bar{q}(l, d_x, t_x), l \in PL_x \tag{6.10}$$

$$\hat{q}(l_x, d_x, t_x) \sim \bar{q}(l, d_x, t_x), l \in TL_x \tag{6.11}$$

For these earlier experiments, only the error indicators are presented to show their performances, changing over fourteen days in three lanes. (The blue line shows the results of the experiment using PLs, while the green line shows results using TLs (lanes that have the same manner of turning).

- ***MAPE***

The results show regular performance for day-of-month. Two sets of lane inputs show opposite performance for different streams or turnings: PLs seems to perform better in left turning (lane 1) and TLs as straight (lane 3). This may due to the fact that flows in left- turning lanes have less similarity than flows in straight lanes. The figures also indicate that approach 2 works better in lane 7 in the south stream than lane 1 or lane 3 in the west stream. This approach provides a smaller percentage error for a large number of streams, which is totally different from the first approach.

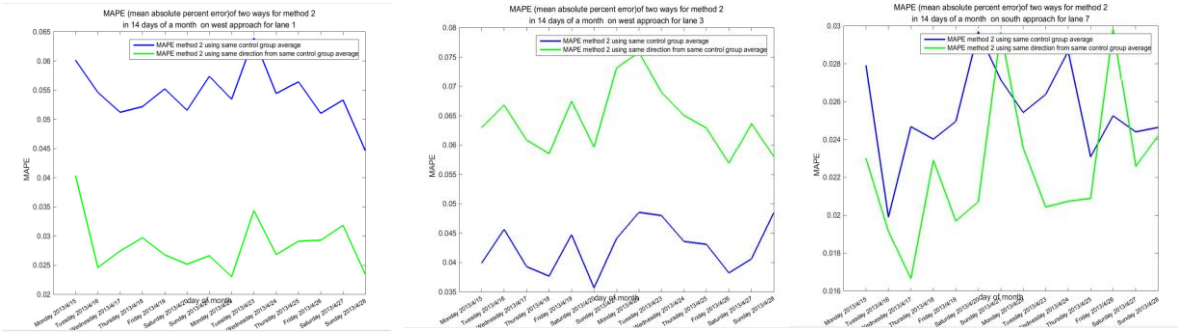


Figure 6-9. MAPE for the approach level, using approach 2 over 14 days per month, based on estimating the missing flow in (a) West stream lane 1 , (b) West stream lane 3, and (c) South stream lane 7.

- ***RSME***

The figures here show that the general output of approach 2 is stable, with little fluctuation over days. However, lane 7 in the south stream, with a higher volume of traffic, still shows a higher absolute error than do the smaller volume lanes.

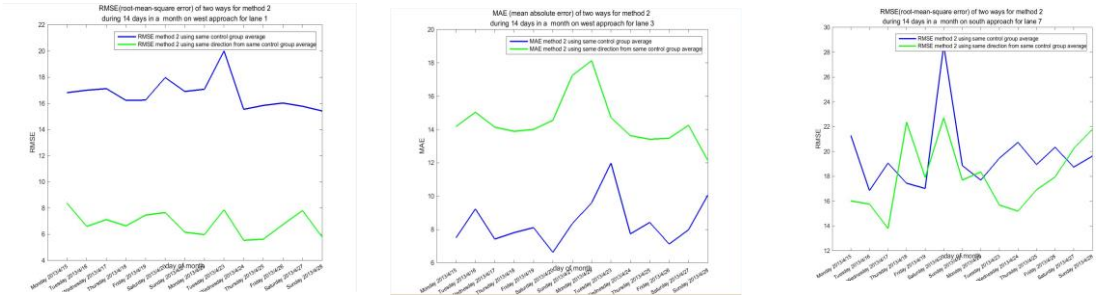


Figure 6-10. RMSE approach level, using approach 2 over 14 days per month, based on an estimation of missing flow on (a) West stream lane 1 , (b) West stream lane 3. (c) South stream lane 7.

The updated algorithm

Given general performances, using primary formulas, the approach is updated, by combining the influence from different sets of lanes. One application involves adding general historical flow ratios between lanes as the weights to be used in upgrading the

method of calculating the contributions from a given set of lanes.

The general performance of the whole junction is presented in Appendix 2. From the tables there, the findings are that the relative errors are stable over different cases (missing flow in different lanes, on different days), and in some cases, the smoothing process does not improve the performance as much as does approach 1.

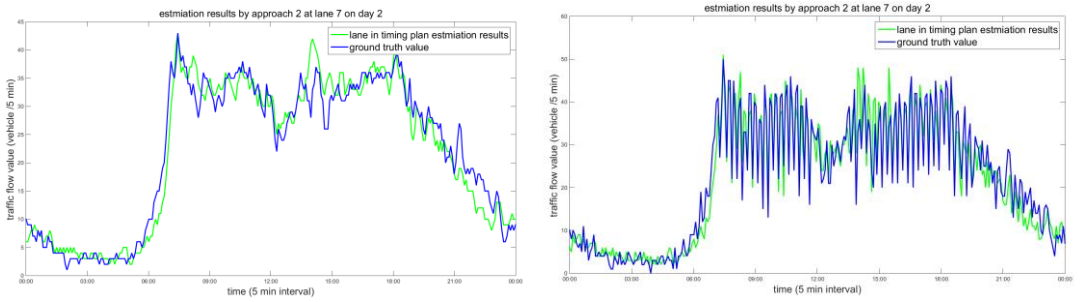
The first formula shows the estimation from another lane from applying the general flow ratio between them. The second formula shows the average of these estimated values for the lanes in the same phase.

$$\hat{q}_l(l_x, d_x, t_x) \sim \left(\frac{\bar{q}(l_x, d_x, t)}{\bar{q}(l, d_x, t)}, t \in T\right) * (\bar{q}(l, d_x, t), t \in T_x) \tag{6.12}$$

$$\hat{q}(l_x, d_x, t_x) \sim \sum_{l \epsilon PL} w_{ph}(l) * \hat{q}_l(l_x, d_x, t_x) \sim \overline{\hat{q}_l(l_x, d_x, t_x)}, l \in PL$$

$$w_{ph}(l) = \begin{cases} 0, & l \notin PL \\ \frac{1}{n_{pl}}, & l \in PL \end{cases}$$

Estimation on the south stream for lane 7, in week 2, on day 2 (April 23, 2013). (The green line shows the estimated values, while the blue line shows actual values).



Error indicator(5min interval)	original data	processed data
MAPE	24.45%	17.18%
RMSE	4.33	3.04

Figure 6-11. Estimation using approach 2 on lane 7, junction 31616, on April 23, 2013 for original data (left) and processed data (right)

The approach 2 lane spatial distribution in a timing plan performed well in capturing even the smallest trend of the traffic flow volume. For the processed data (smoothed flow volume), the performance is almost the same as in the first approach. It is better than the first approach in the original data (original interrupted values), if only in terms of the error indicators.

For approach 2 (A2), lane spatial distribution, the general results fluctuate less than approach 1.1, and the results from primary implementation are around 10% and 40%. The upgraded implementation shows a performance of around 25% on original data and around 15% on processed data in this case. Some of the gains include the fact that the flows distributed on lanes with different turning groups are of various regularities.

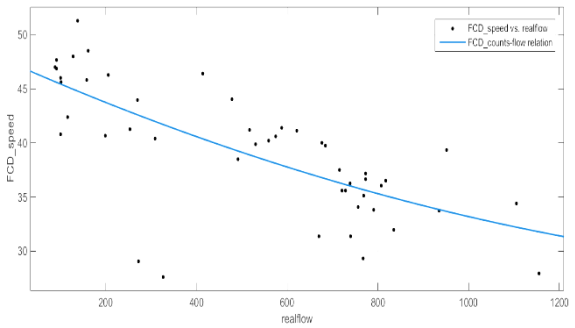
For instance, the flows in left-turning lanes have less similarity than that in straight lanes. Also, this approach provides a smaller percentage error for a larger number of streams, which is totally different from the historical pattern approach. Comparing the difference between performance on original and processed data, this shows less difference. It may be concluded that this approach is more suitable than approach 1.1 under situations in which there are large fluctuations of flows over days-in-a-week.

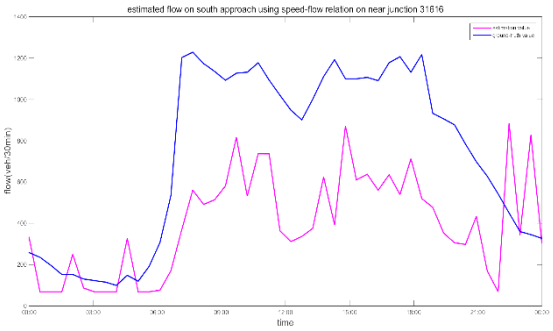
6.4. Cases for approach 3 FCD - flow data fusion

FCD can only be as precise on a road segment level, so this approach acts on a stream level rather than an individual lane level. Speeds and counts of FCD on April 23, 2013 at this junction are calculated. The flows on the segment are the sum of flows from lanes in this same approaching direction. The relations are presented in Appendix 3.

6.4.1. FCD Speed-loop flow

Estimation of the south stream for week 2, day 2 (April 23, 2013). In the curve in Figure 6-12, the x-axis represents flow and the y-axis speed. In the results, the pink line shows the estimated values while the blues line shows actual values.





Error indicator(30min interval)	original data
MAPE	55.10%
RMSE	455.91

Figure 6-12. Fitting curve of outbound speed and flow for south stream on junction 31616 (left), and estimation using approach 3.1 on outbound south stream on junction 31616 for original data (right)

Some more fitting results are presented in appendix 3. These figures show that speed-flow relations do exist; however, the relations are not unified. The flows are underestimated using this approach, with a relative error of 55.10%, which is not a satisfactory output. The possible reasons are assumed:

1. The flow-speed relations are hard to obtain at junctions, since the traffic volumes are composed of several interrupted streams, and are influenced

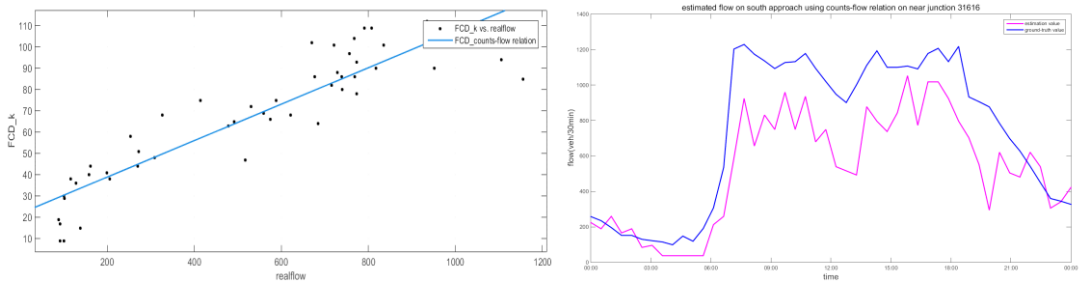
largely by the signal control. Although, in some streams, the fitting curves of speed and flow show the same shape as the hypothetical speed-flow curve, though this relation is not obvious on some streams.

2. The speed-flow relations are not sharp or distinguished enough to capture the flow values precisely, and thus do not easily provide precise outputs. It is easy to understand the placid slopes here: A stop line can lead to many lower speed or even static speed records, which decrease the average speed calculated from each time interval; this lowers the slope of the whole fitting curve.

However, this approach is theoretically sound, along with the data fusing concept. It also captures the major trend of relation between flows and speeds. Therefore, it can at least be used to provide reference values, which may be not so accurate temporarily, but may be better than none.

6.4.2. FCD Counts-loop flow

The estimation using this algorithm is conducted on the south stream in week 2, day 2 (April 23, 2013). In the fitting curve shown in Figure 6-13, the flows are shown on the x-axis, and the counts on the y-axis. In the results, the pink line shows the estimated values, while the blues line shows actual values.



Error indicator(30min interval)	original data
MAPE	32.24%
RMSE	273.06

Figure 6-13. Fitting curve of count and flow outbound of south stream on junction 31616 (left), and estimation using approach 3.2 outbound of south stream junction 31616, April 23 2013, for original data (right).

The FCD counts-flow relation shows a trend with an obvious direct ratio. The relations in other areas are shown in appendix 3, and these relations are quite similar. In the estimation, as in the speed-flow approach, the flows are also underestimated, with a relative error of 32.24%. This result is better than that using the speed-flow relation. Although the penetration rate changes during the time of day, the FCD counts still represent one part of the total traffic flow. Therefore, the increase in FCD counts may imply an increase in the total flow, generally speaking.

By way of a short conclusion, Approach 3 (A3) contains two algorithms (sub-approaches), involving estimations made using the FCD speed-flow relation and the

FCD counts-flow relation. The relations between speed and flow vary in different approaching directions, even at the same junction; while the relations between count and flow are relatively stable, almost with a direct ratio.

6.5. Cases for approach 4 multiple linear regression

For Approach 4 (A4), using a multiple linear regression (MLR), as stated above, the choices of input range (relevance) and analysis interval are the key factors influencing performance. Also, the performance over the course of a day is analyzed, and this shows how the performance can changes during a single day for a certain algorithm.

To find out the influence from the factor of input range, the analysis interval is fixed as 24h. First, all the available observations at the junction are used as the input dataset. Secondly, the range is narrowed, such that only detectors from one approaching direction are selected. To test the influence from the factor of analysis interval, the input range is fixed, detectors from one approaching direction are selected, and four different analysis intervals are tested. To test the performance over several periods, all cases are selected, and the evaluation interval is changed from one day to one hour.

For these tests, the inputs last a week; for all observations, the parameters are for one week and they are applied to another. The results for all of the cases are in Appendix 4. Some cases are presented here for analysis.

Scenario 1 Influence of inputs range

The estimation results obtained from the regression on all other detectors seems rough. Some trends and peaks are totally different, especially during the daytime. It is understandable: During the daytime, the change in traffic volumes for each lane from each stream are frequent, and contributions from other observations can be quite unreliable.

Narrowing down the input range has made the relative errors smaller, from 42.68% to 25.67% for original data, and from 26.81% to 19.02% for processed data. The absolute errors have also declined. These results give evidence that more inputs for MLR do not lead to more reliable results. The observations that are relevant play an important role.

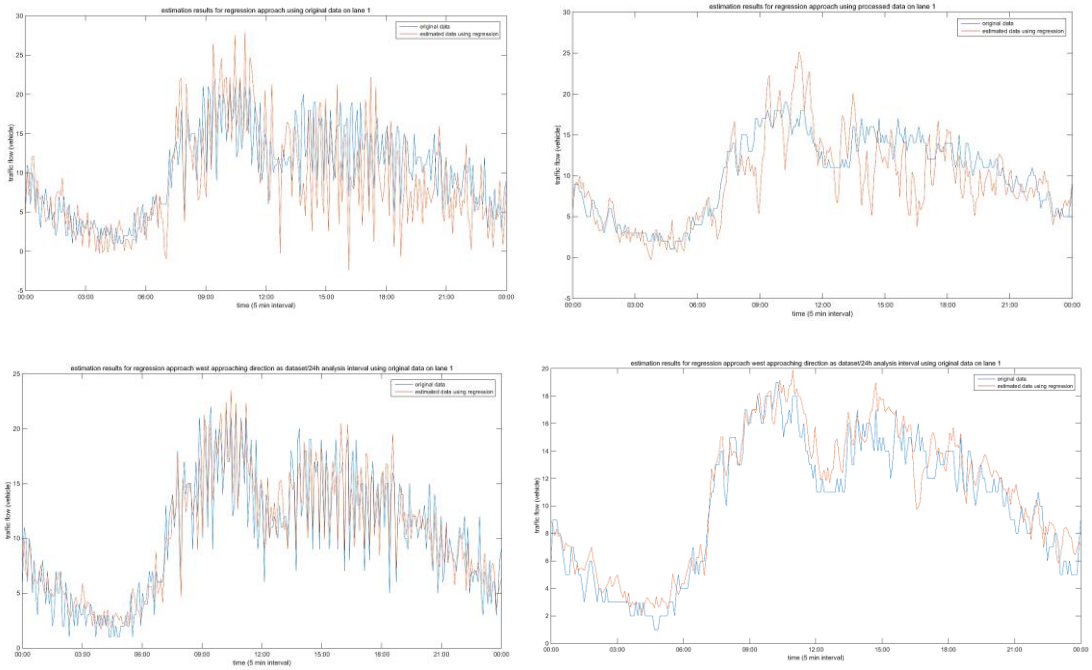


Figure 6-14. Estimation of missing flow in lane 1 on April 22, 2013, using the multiple linear regression approach: use the original data case (left) and processed data case (right). (Top: the whole dataset is from one approaching stream, down: dataset from the West stream. Analysis interval: 24h).

Table 6-1. Error indicators for approach 4, lane 1 April 22, 2013.

	Inputs	original data	processed data
MAPE	the whole junction	42.68%	26.81%
	one steam	25.67%	19.02%
RMSE	the whole junction	4.37	3.38
	one steam	2.01	1.65

Scenario 2 Influence of analysis interval

According to section 4.2.4, the minimum analysis interval to ensure a reliable regression can be as short as 4 hours. The analysis intervals of 24 hours, 12 hours, 8 hours, and 4 hours are tested.

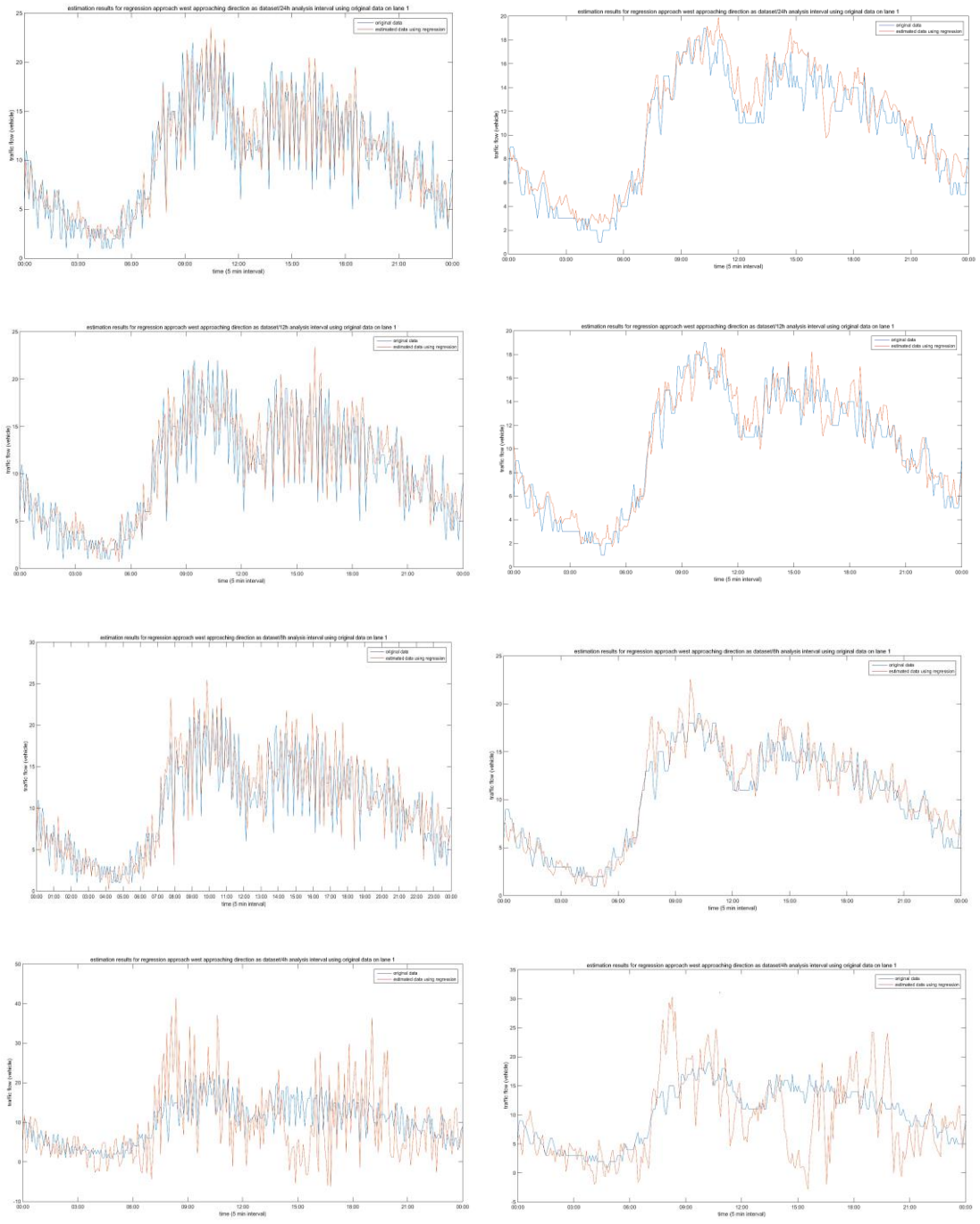


Figure 6-15. Estimation of missing flow in lane 1, on April 22 2013, using the multiple linear regression approach: use original data (left) and processed data (right). (The whole dataset is from the West approaching stream, and the analysis interval is from top to bottom: 24h, 12h, 8h, 4h).

Table 6-2. Error indicators for approach 4, lane 1, on April 22, 2013; validation interval: 1 hour

	Analysis interval	original data	processed data
MAPE	24 h	25.67%	19.02%
	12 h	30.70%	14.93%
	8h	31.21%	16.20%
	4h	71.55%	43.65%
RMSE	24 h	2.01	1.65
	12 h	2.54	1.36
	8h	2.87	1.66
	4h	7.86	5.54

The results, with different analysis intervals, show similar trends. Results with analysis interval 24h and 12h do not show much difference in performance. Results using 4h provide a larger error (MAPE of more than 70%) than other cases. This shows that the suitable analysis interval is among 8, 12h, and 24h.

Scenario 3 Performance over period

Figure 6-16 gives the MAPE during each period of a single day for several applications under different analysis intervals for approach 4; approaches 1 and 2 are also shown for comparison. Previous approaches seem steady every time. MLR approaches show larger errors at midnight and smaller errors during daytime.

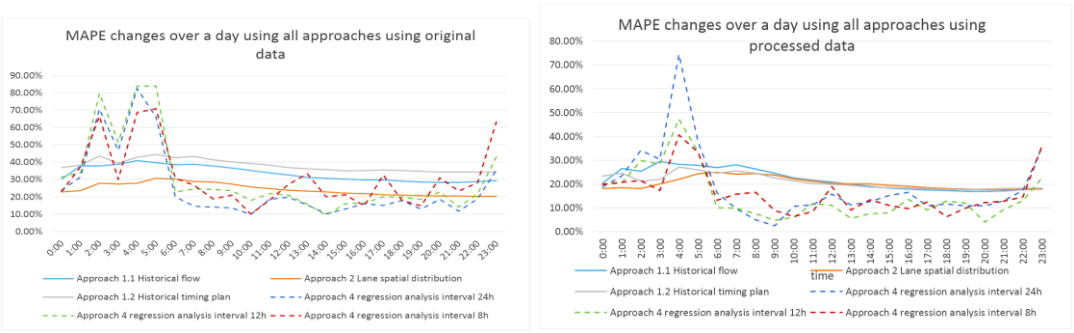


Figure 6-16. Comparison of each approach on MAPE, on lane 1, April 21, 2013 ; validation interval: 1 hour; using original and processed data.

In conclusion, for the input range, relevant inputs contribute to accurate estimation results. For the analysis interval, there are different performances under different analysis intervals. No matter for which analysis inputs, the accuracy using this approach increases with the increase in the flow amount concerned. It acts better at peak hours with less than 10% errors, and worse during the night, with 40% errors or more.

6.6. Cases for integrated method 1: Iteration

This method combines approaches 1 and 2, using an iterative process. The iteration has been described in section 4.3.1. The specific formulas are as follows:

Initial input from historical flow pattern (Approach 1.1):

$$\hat{q}_1(l_x, d_x, t_x) \sim \{w_{DOW} * \bar{q}(l, d_{DOW_x}, t), w_{WD} * \bar{q}(l, d_{WD_x}, t), w_D * \bar{q}(l, d_{AD}, t), t \in T_x\}, l \in L \quad (6.13)$$

Initial input from spatial distribution in a timing plan (Approach 2):

$$\hat{q}_2(l_x, d_x, t_x) \sim \sum_{l \epsilon L} w_c(l) * w_{tr}(l) * w_{pl}(l) \frac{1}{n_{pl}} \left(\frac{\bar{q}(l_x, d, t)}{\bar{q}(l, d, t)}, t \in T \right) * (\bar{q}(l, d, t), t \in T_x), d \in D \quad (6.14)$$

Two scenarios are considered, according to missing data type: long-term missing (flow is missing for the whole day) and short-term missing (flow is missing for some period). Each scenario presents the results of estimation using original or smoothed data and corresponding iteration times before convergence. Some representative cases are presented: The estimation on Week 1, Day 1 (April 15, 2013) for Lane 7, and on Week 2, Day 2 (April 23, 2013) for Lane 7. Finally, a comparison of the method for individual approaches is made.

Scenario 1: missing for a long term

A convergence has been reached for each data point. For the majority of individual values, the iteration time is 7, the highest is 9, and the lowest 5 times.

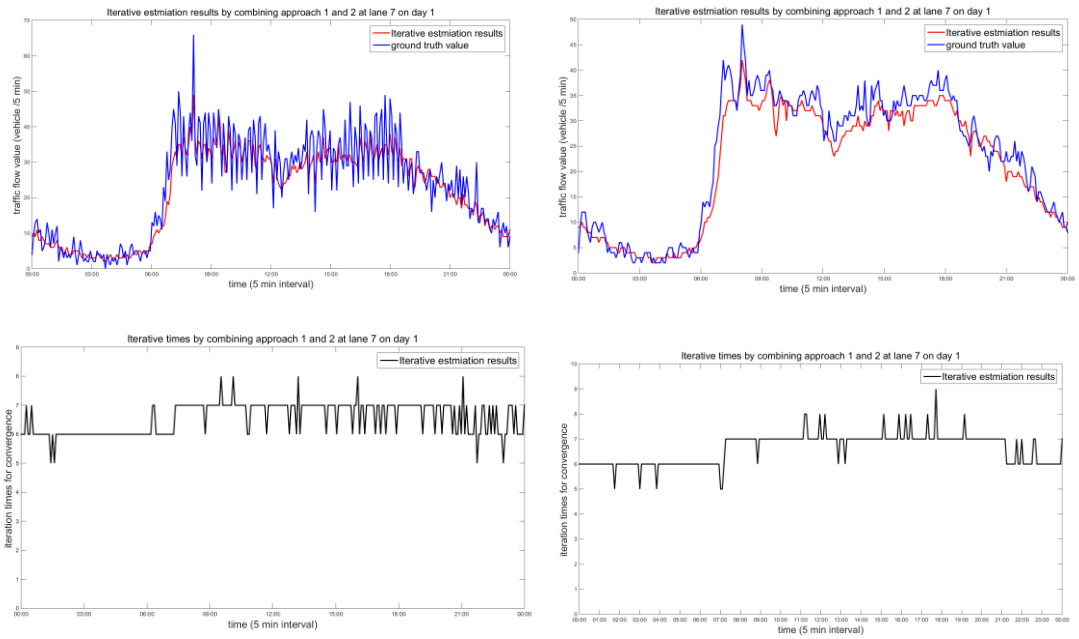


Figure 6-17. Iteration results for long-term missing (up: flow compared with actual detected flow, down: iteration times before convergence), lane 7, week 1, day 1 (April 15, 2013), original data (left) and processed data (right).

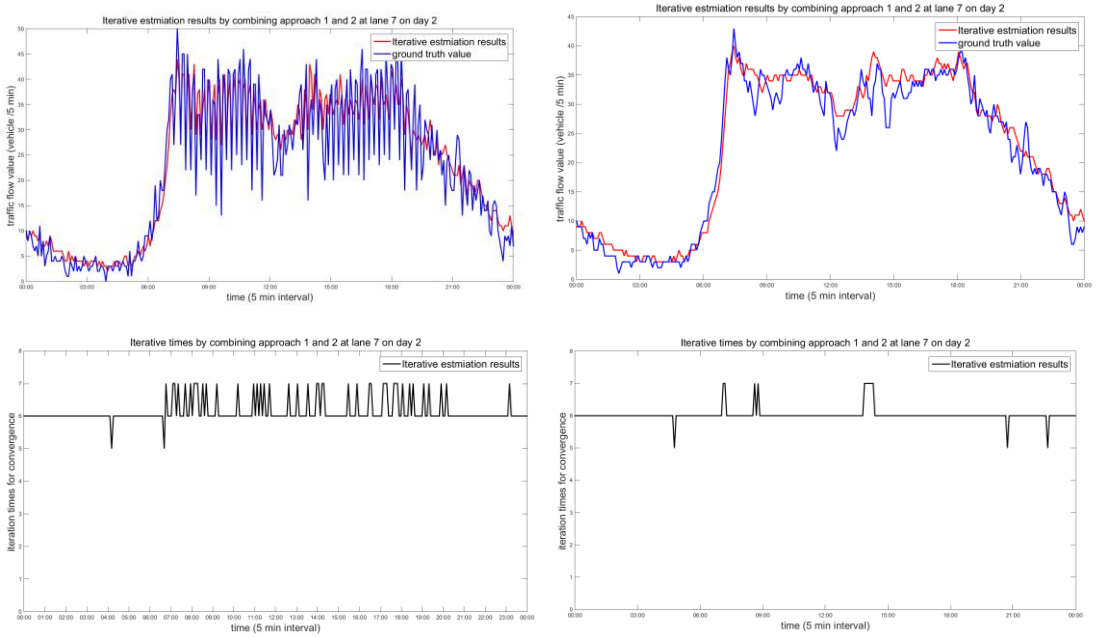


Figure 6-18. Iteration results for long-term missing (up: flow compared with actual detected flow, down: iteration times before convergence), lane 7, week 2, day 2 (April 23, 2013), original data (left) and processed data (right).

Table 6-3. Error indicators for iterative estimation for long-term missing, on lane 7, April 15 and 23, 2013.

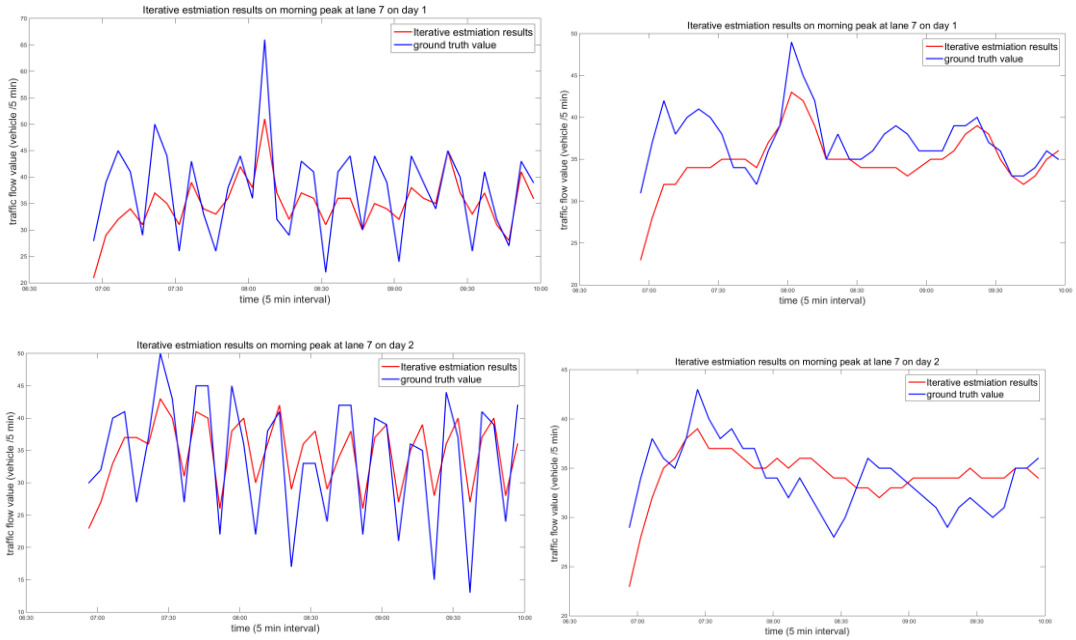
	Day	original data	processed data
MAPE	Week 1 day1	25.16%	14.03%
	Week 2 day2	27.52%	15.84%
RMSE	Week 1 day1	5.18	3.13
	Week 2 day2	4.93	2.52

Scenario 2: missing for a short term

During the peaks in both morning and afternoon, estimated values follow closely the actual detected values for original and processed data.

- ***Morning peak 7:00-10:00***

During the peak in the morning, estimated values vary closely with actual detected values, as with the processed data case.

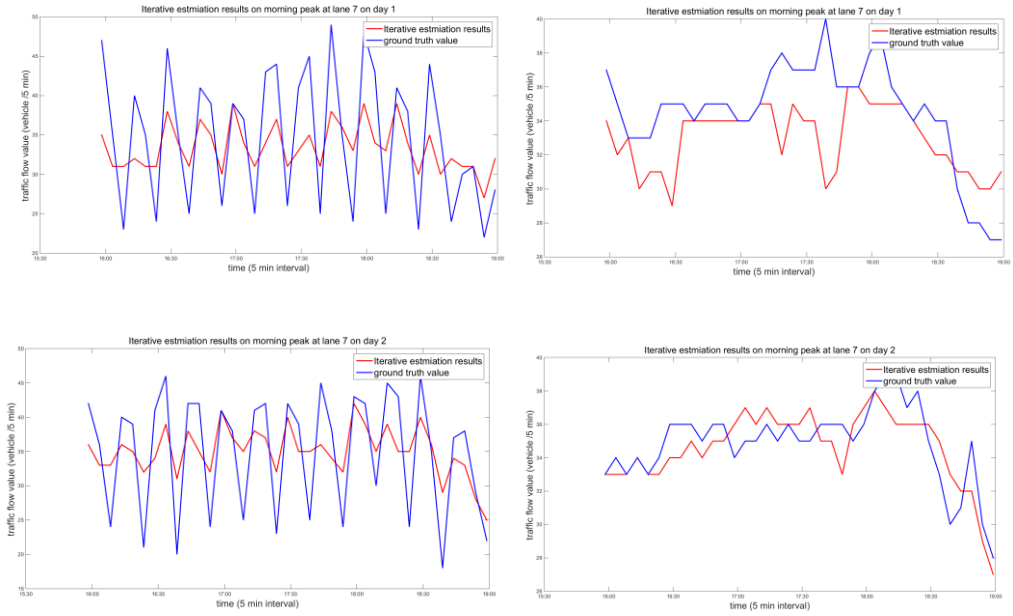


	Day	original data	processed data
MAPE	Week 1 day1	14.17%	7.55%
	Week 2 day2	19.66%	8.09%
RMSE	Week 1 day1	6.40	3.94
	Week 2 day2	6.09	3.18

Figure 6-19. Iterative estimation for short-term missing morning peak, 7:00-10:00, on lane 7, April 15 and 23, 2013; original data (left) and processed data (right).

- ***Afternoon peak 16:00-19:00***

During the peak in the afternoon, estimated values follow closely the actual detected values, with less variance. For processed data case (smoothed), the estimated value corresponds with the same trends as the actual detected values.



	Day	original data	processed data
MAPE	Week 1 day1	17.69%	19.42%
	Week 2 day2	6.46%	3.64%
RMSE	Week 1 day1	6.60	3.08
	Week 2 day2	6.54	1.53

Figure 6-20. Iterative estimation for short-term missing afternoon peak, 16:00-19:00, on lane 7, April 15 and 23, 2013; original data (left) and processed data (right).

The estimation results for the short-term missing type are better than that for the long-term type. For lower resolution, as shown in Appendix 5, such as for the 15-minute and 30-minute intervals, the performances are more stable and with fewer relative errors.

Comparison: Integration 1 compared with approaches 1 and 2

The green color shows the initial results from approach 1, while yellow shows approach 2, and the red line shows the iterative results. When looking at each time stamp, the majority of the iterative estimation values are in-between individual approaches. However, not all of the final values are in-between the values from individual approaches: this phenomenon can be explained by the fact that results from the individual approaches are all ones of "initial" estimation without updating from the dimension. During the process of iteration, the missing flow has been updated gradually by the stepwise estimations, and these newly updated values will contribute to the weights or ratios considered in both approaches. Thus, the estimation from the next iteration may be quite different from the initial estimate.

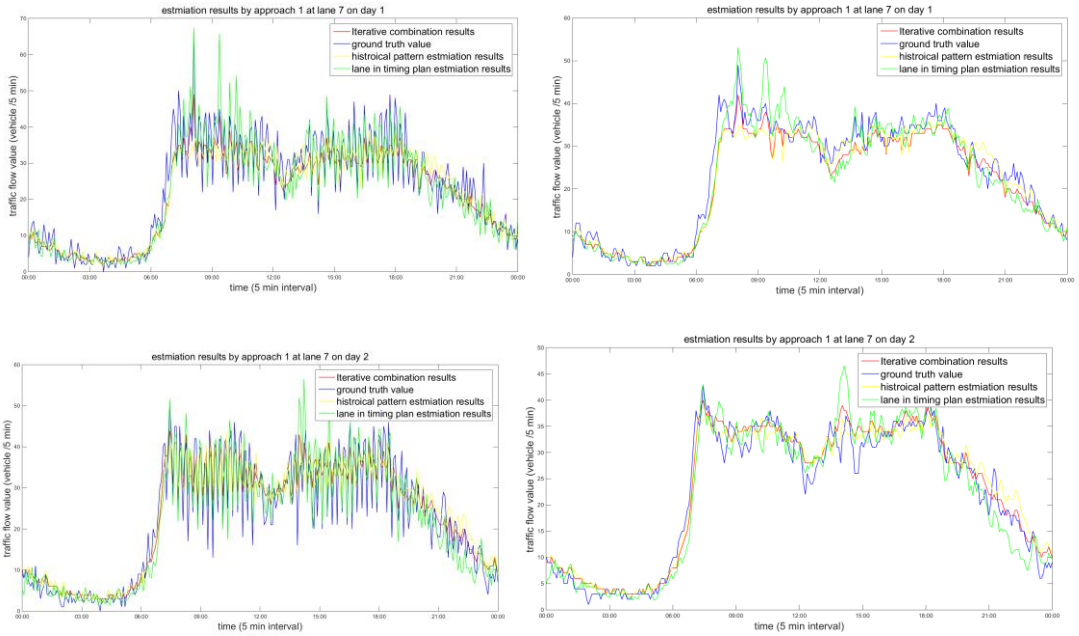


Figure 6-21. Iterative estimation and approaches 1 and 2 for the long-term missing, in lane 7, on April 15 and 23, 2013; with original data (left) and processed data (right).

Table 6-4. Error indicators for iterative estimation, and approaches 1 and 2 for long-term missing, on lane 7, April 15 and 23, 2013.

			original data			processed data
Error indicator	A1	A 2	Iterative results	A1	A2	Iterative results
MAPE	29.97%	24.83%	25.16%	14.57%	15.96%	14.03%
RMSE	6.66	4.95	5.18	3.34	3.66	3.13
MAPE	34.70%	27.12%	27.52%	17.80%	18.40%	15.84%
RMSE	6.90	5.03	4.93	2.75	3.61	2.52

When one of the approaches does not perform well, the iterative tend to be closer to the better-performing one. When the results from two approaches are close, results from integration using iteration can provide a better estimation than any one of them. Since there is no idea when an approach will perform better than another one, using an iterative one is a reliable and safe solution.

6.7. Cases for integrated method 2: Advanced MLR

This method is based on approach 4 MLR, considering the relevant information from approach 1 and 2, used to support the selection of the inputs. The formula is from section 4.3.2, and two sets of relevant inputs are selected: $q(l,d,t)$, $l \in PL$, and $q(l,d,t), d \in DOW_x$. The formula used here is:

:

$$\hat{q}(l_x, d_x, t_x) \sim f_{regress}(q(l, d, t), \beta), l \in PL, d \in DOW_x, t \in T \qquad (6.15)$$

Two scenarios are set: missing flow for the long- and short-term. For each scenario, the cases are still on the south stream for lane 7, on April 15 and 23, 2013. The regression analysis interval is 24h, and the dataset is the approaching stream wherein the lane lay. (The dashed red line shows the estimated values, while the blue line shows actual values).

Scenario 1: missing over a long term

The results of advance MLR are slightly better than using the iteration. The relative error is around 25% for original data and 15% for processed data.

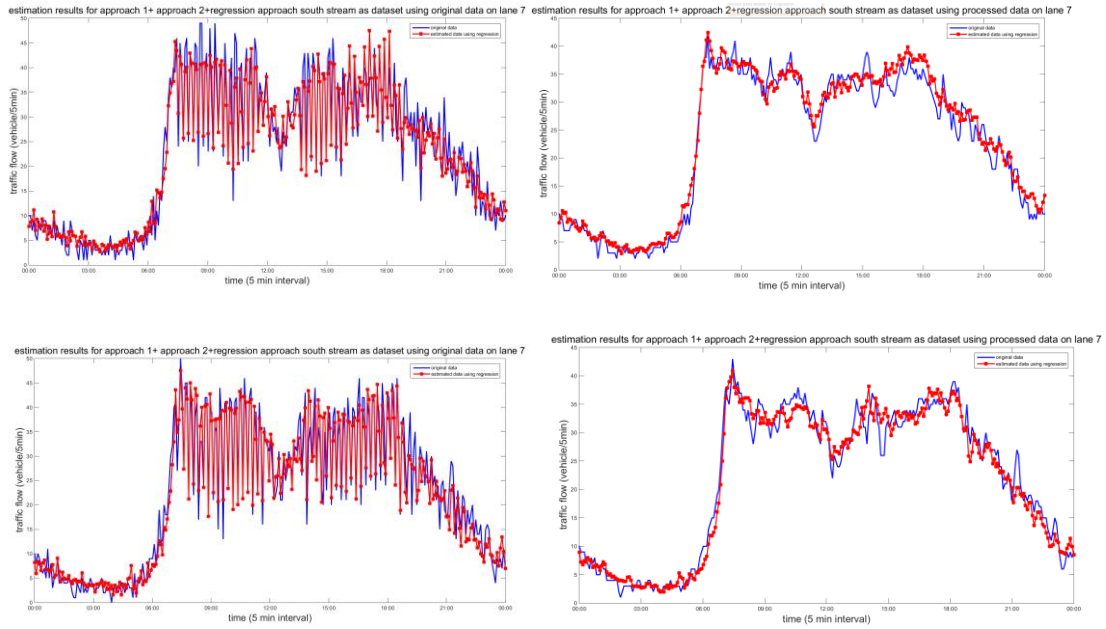


Figure 6-22. Estimation using integration 2 for original data (left) and processed data (right) on lane 7, junction 31616, on April 15 and 23, 2013.

Table 6-5. Error indicators using integration 2 on lane 7, junction 31616, on April 15 and 23, 2013.

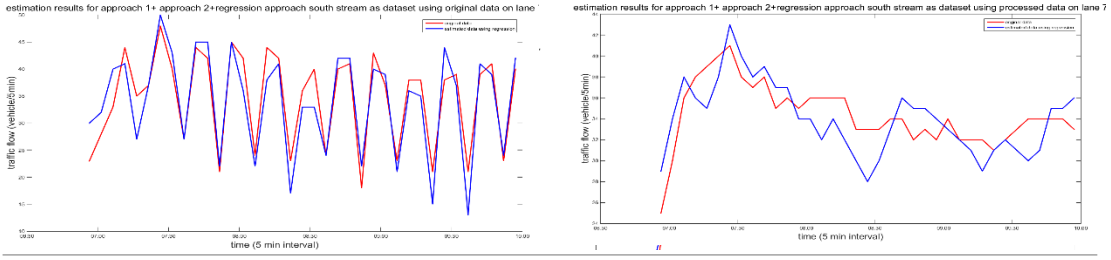
	Day	original data	processed data
MAPE	Week 1 day1	24.90%	16.16%
	Week 2 day2	21.68%	13.16%
RMSE	Week 1 day1	3.30	2.62
	Week 2 day2	3.53	2.20

Scenario 2: missing over a short term

The relative errors in short term missing flow estimation are around 10% of original

data and 5% for processing data, which is lower than the long-term, so it is for RMSE. The advanced MLR shows better performance for missing on short -term than long-term.

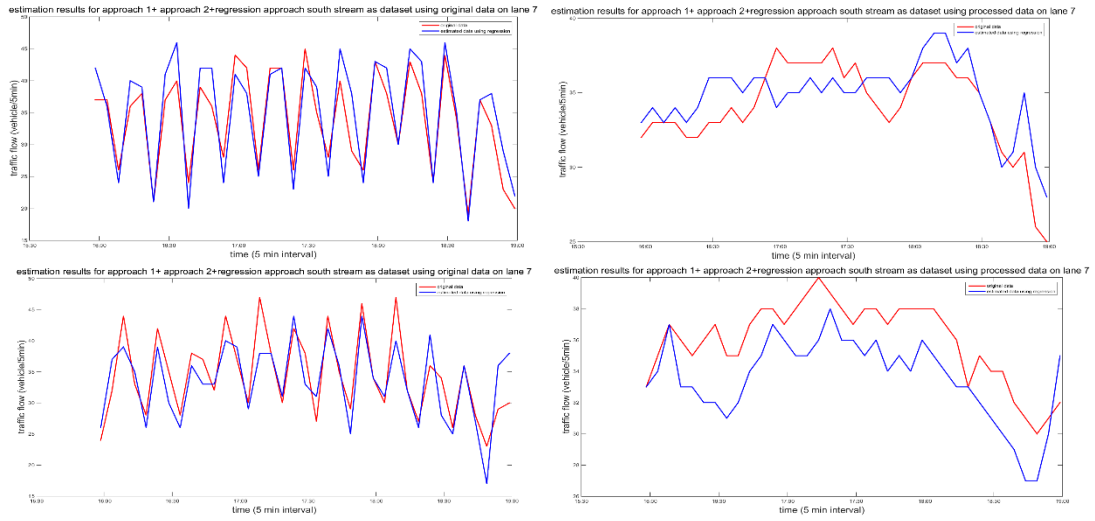
- ***Morning peak 7:00-10:00***



	Day	original data	processed data
MAPE	Week 1 day1	11.47%	6.59%
	Week 2 day2	9.58%	4.84%
RMSE	Week 1 day1	3.95	2.47
	Week 2 day2	4.15	1.85

Figure 6-23. Estimation using integration 2 for original data (left) and processed data (right) on lane 7, junction 31616, on April 15 and 23, 2013, during morning peak, 7:00-10:00.

- ***Afternoon peak 16:00-19:00***



	Day	original data	processed data
MAPE	Week 1 day1	10.55%	5.61%
	Week 2 day2	6.96%	4.79%
RMSE	Week 1 day1	3.47	2.11
	Week 2 day2	2.92	2.02

Figure 6-24. Estimation using integration 2 for original data (left) and processed data (right) on lane 7, junction 31616, on April 15 and 23, 2013, during afternoon peak, 16:00-19:00.

In conclusion, an advanced MLR shows a low error in both long-term and short-term

MLR, and performs better in the short-term. It captures the trend of the actual data well, which shows the advantage of considering similar or nearby lanes. Thus, this method has certain advantages since it selects inputs by considering both approaches 1 and 2; that is to say, it uses the observations from the most relevant days and lanes.

6.8. Cases for validation

From sections 6.2 to 6.5, test results are provided, along with the calibration of four individual approaches. In sections 6.6 and 6.7, two integrated methods are tested. Individual approaches 1, 2, and 4 are widely applied to cases on different lanes and days (see Appendices 1, 2, and 4). Therefore, these results can be seen as a kind of validation. In approach 3, the relations are irregular in a different stream (see Appendix 3), so there is no suitable means of validation.

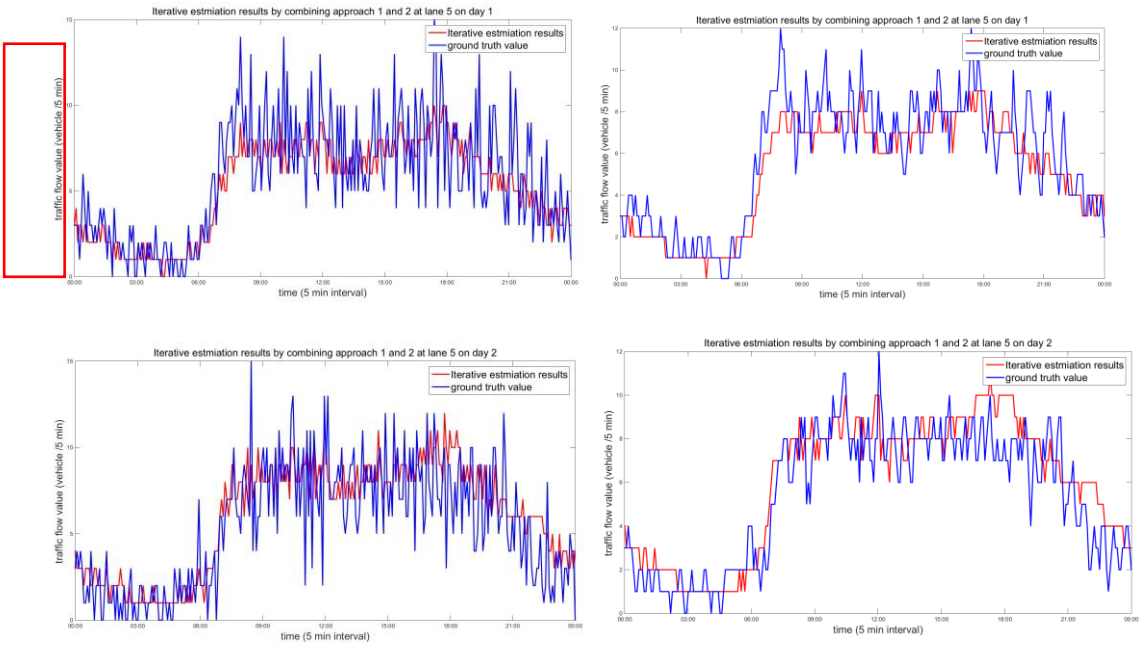
In this part, the main task is to apply two integrated methods (Iteration and advanced MLR) to another junction (junction 31617) by reproducing the similar experiments in sections 6.6 and 6.7.

From the experiment setup in Figure 6-2 , the layout of junction 31617 is similar to 31616. Lane 5 at junction 31617 expresses a similar position to lane 7 at junction 31616 (on the south-approaching stream, first left-turning lane). This section thus applies two methods to lane 5 to see if they still work. Other experimental set-ups are the same as the test cases in sections 6.6 and 6.7. Two scenarios including the long-term and short-term missing. Each scenario uses original and processed (smoothed) data. Two sets of missing flow are used for evaluation: April 15, 2013 (week 1, day 1) and April 23, 2013 (week 2, day 2) for lane 5.

6.8.1. Validation of iteration method

Scenario 1: missing over a long term

Inspecting the results for long-term missing flow, the relative errors are around 30% for original data and 25% for processed data, which is larger than the errors in section 6.6. However, the absolute errors are smaller: around 2 for original raw data and around 1.5 for processed data. The explanation is that the total traffic volume in this new lane at this new junction is much smaller than in the lane at the junction in previous cases.



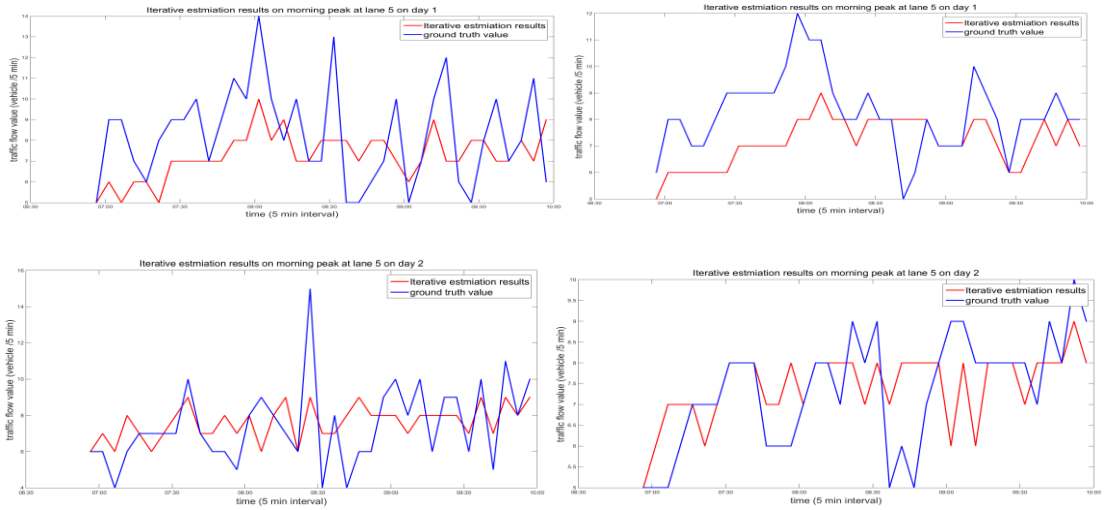
	Day	original data	processed data
MAPE	Week 1 day1	29.22%	18.56%
	Week 2 day2	38.81%	29.79%
RMSE	Week 1 day1	2.16	1.40
	Week 2 day2	2.11	1.44

Figure 6-25. Iterative estimation for long-term missing, on lane 5, April 15 and 23, 2013; original data (left) and processed data (right).

Scenario 2: missing over a short term

For short-term cases during the morning and afternoon peaks, the relative errors are around 20% to 30% for original raw data, and 15% for processed data. Similarly to the long-term case, compared to the test cases in section 6.6, the relative errors are larger and the absolute errors are smaller.

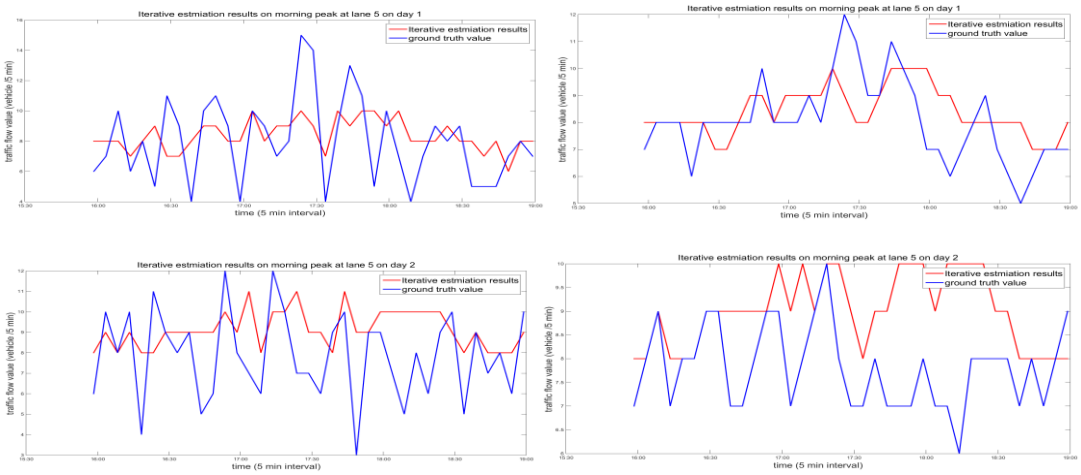
- ***Morning peak 7:00-10:00***



	Day	original data	processed data
MAPE	Week 1 day1	23.23%	16.08%
	Week 2 day2	22.04%	13.05%
RMSE	Week 1 day1	2.31	1.72
	Week 2 day2	1.91	1.20

Figure 6-26. Iterative estimation for short-term missing morning peak, 7:00-10:00, in lane 5, April 15 and 23, 2013; original data (left) and processed data (right).

- ***Afternoon peak 16:00-19:00***



	Day	original data	processed data
MAPE	Week 1 day1	32.43%	13.56%
	Week 2 day2	32.45%	17.17%
RMSE	Week 1 day1	2.57	1.40
	Week 2 day2	2.46	1.63

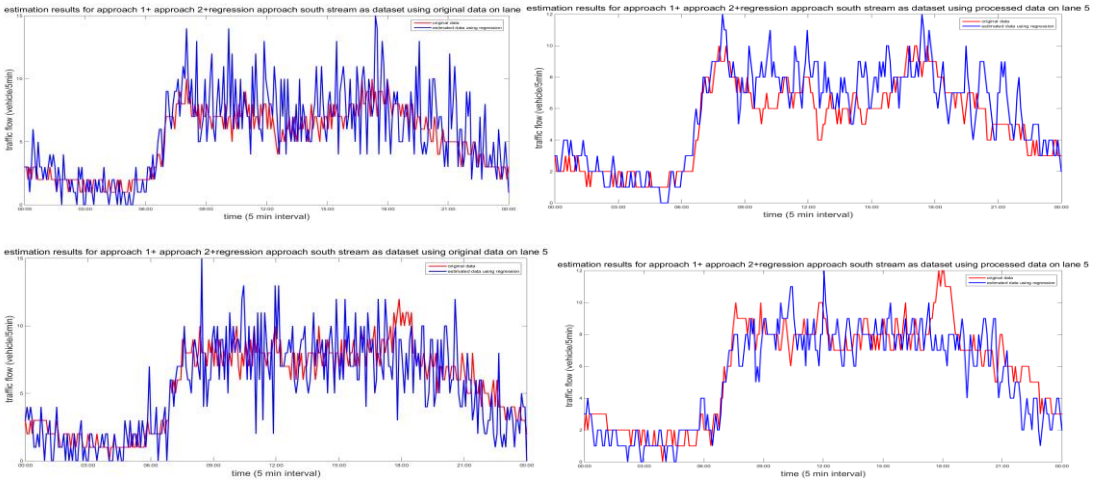
Figure 6-27. Iterative estimation for short-term missing afternoon peak, 16:00-19:00, in lane 5, April 15 and 23, 2013; original data (left) and processed data (right).

In conclusion, when applying the method in a lane at another junction, the results (differences between the new case and the previous cases) are within an acceptable range according to the setup of validation. Due to lower traffic volume, the new application turns out larger relative errors and smaller absolute errors. Considering these two error indicators together, the difference in performance between the test and validation cases is reasonable and acceptable. Thus, the integrated method using iteration is validated.

6.8.2. Validation of advanced MLR

If the results for missing flow over the long term are inspected, the relative error is around 35% for original raw data and 30% for processed data, which is larger than errors in section 6.6. However, the absolute error is still smaller—around 2.5 for original raw data and 1.5 for processed data. Both indicators are worse than iteration.

Scenario 1: missing over a long term



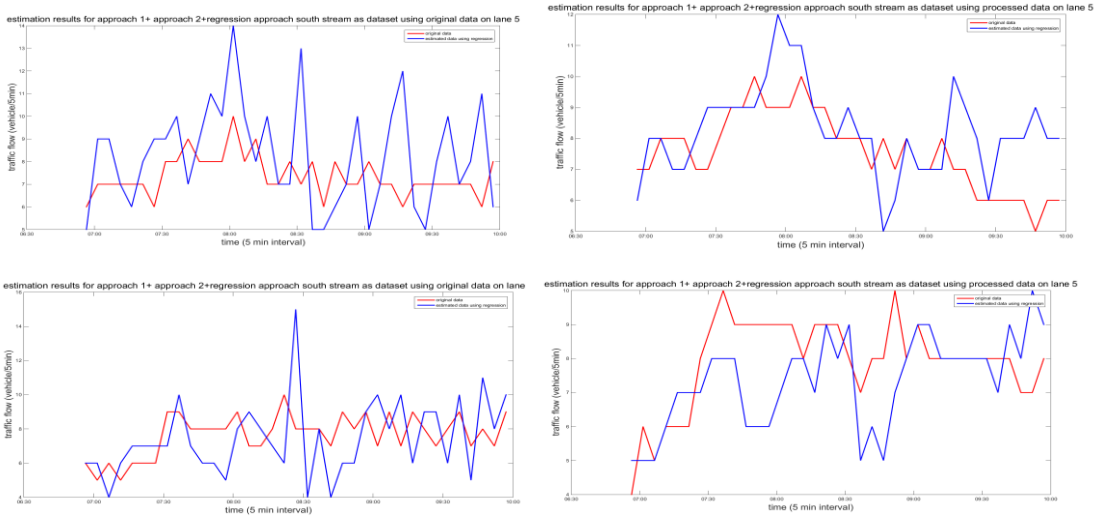
	Day	original data	processed data
MAPE	Week 1 day1	36.24%	24.01%
	Week 2 day2	49.14%	33.93%
RMSE	Week 1 day1	2.51	1.73
	Week 2 day2	2.29	1.69

Figure 6-28. Iterative estimation for long-term missing, on lane 5, April 15 and 23, 2013; original data (left) and processed data (right).

Scenario 2: missing for a short term

Over the short-term on morning peak and afternoon peak, the relative errors are around 25% to 40% for original raw data and 20% for processed data. Compared to results in long-term cases, they are fewer. Compared to the test cases in section 6.6, the relative errors are larger and the absolute errors smaller. The results are still less promising than the results of iteration cases.

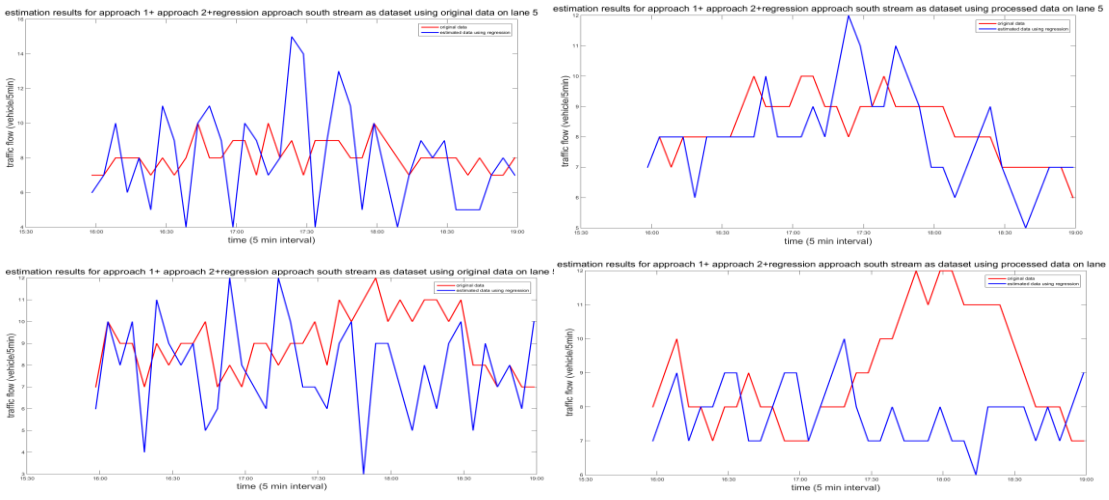
- ***Morning peak 7:00-10:00***



	Day	original data	processed data
MAPE	Week 1 day1	39.68%	24.71%
	Week 2 day2	25.55%	13.16%
RMSE	Week 1 day1	3.01	2.04
	Week 2 day2	2.16	1.27

Figure 6-29. Iterative estimation for long-term missing, in lane 5, on April 15 and 23, 2013; original data (left) and processed data (right).

- ***Afternoon peak 16:00-19:00***



	Day	original data	processed data
MAPE	Week 1 day1	32.09%	22.64%
	Week 2 day2	37.14%	20.35%
RMSE	Week 1 day1	4.21	2.87
	Week 2 day2	3.15	1.93

Figure 6-30. Iterative estimation for long-term missing, in lane 5, on April 15 and 23, 2013; original data (left) and processed data(right).

The selected comparison lane at the new junction has a lower quantity of traffic flow. For both methods, the MAPE is larger than that of calibration cases, but these indicators are still around 20% to 30%, which is acceptable. Since the new lane at the new junction has smaller traffic volume, these poorer results have also verified the conclusion that, for MLR, a larger amount of the value provides higher accuracy.

In conclusion, and taking both MAPE and RMSE into consideration, the difference in error indicators in the validation cases (similar lanes from a new junction on the same days) is within the acceptable range. Advanced MLR, together with Iteration, is validated, and they can be applied to other lanes or other junctions to make the estimation.

6.9. Conclusion for the chapter

The approaches and methods are tested and evaluated in this chapter from many angles in multiple scenarios.

Some of the results are validated, Approaches 1.1, 2, and 4. As shown in appendices 1, 2 and 4, their performance fluctuates over different days on different lanes. A grand average is calculated for each approach in Table 6-6 (for an approach to 4 MLR, the analysis interval is 24 h). In this table, the historical pattern (A1) is best for processed data, while MLR (A4) performs best with original data. Lane spatial distribution (A2) has the smallest error when considering both situations.

Some approaches or sub-approaches still remain to be improved. They are Approaches 1.2 and 3. Approach 1.2, which uses an historical timing-flow pattern, does not work when the timing is unchanged. Approach 3.1 using the speed/flow relation, which provides irregular relations over different streams (see appendix 3). For Approach 3.2, though it gives an obvious direct ratio between FCD counts and total flow, and the results in the case are better than with Approach 3.1, it does not have theoretical support, and the penetration rate changes over time (shown in section 3.4.2) constitute an underlying problem.

Table 6-6. Error indicators for three individual approaches, grand average of MAPE

	Sub approach	original data	processed data
Historical pattern (A1)	Historical flow pattern	41.25%	24.63%
Lane spatial distribution(A2)	/	38.01%	27.48%
MLR(A4)	All observations	55.03%	45.44%
	One stream	37.80%	27.93%

As for two integrated methods, they both are promising of results, and their performance is close to each other.

Integrated method 1 (I1) iteration show errors in the estimation at around 25% (for original data) and 15% (for processed data) and is better than any single application of approaches 1 or 2. On the one hand, whenever one of the individual approaches fails to perform properly, the iteration process can revise the result gradually and reliably to a certain degree. On the other hand, the iteration provides better results when the rules on both sides are normal. Therefore, using the iterative methods is a reliable, safe solution. The only drawback is the high computation costs.

After calibration, it is suggested that I2 uses the specific approaching stream as inputs and 24h as analysis interval. This method in fact shows the fusion of relevant information to a regression model, which involves better performance than just applying single MLR by using excessively wide inputs. It performs slightly better than I1, with error rates of nearly 20% (for original data) and 10% (for processed data) errors when total traffic volume is relatively large (as seen in the test cases in section 6.7). However, it performs worse than I1, with error rates of more than 30% (for original data) and 25% (for processed data) when dealing with smaller traffic volumes (as seen in the validation cases presented in section 6.8).

Table 6-7. Error indicators for I1 and I2 in lane 7 at junction 31616, on April 23, 2013.

Error indicator	MAPE		RMSE	
	original data	processed data	original data	processed data
Iteration (I1)	27.52%	15.84%	4.93	2.52
Advanced MLR(I2)	21.68%	13.16%	3.53	2.20

7. Conclusions

In this chapter, comments are made on each approach and method, and a comparison is made among them. The best methods for estimating missing flow at the urban junction are determined according to the specific situations. The methods are recommended for practical use in estimating flows at urban junctions under stable traffic conditions according to the situations referring to the criteria mentioned in this thesis.

Comments on methods

A comparison is made of all of the individual approaches as well as integrated methods. All of the major criteria observed from tests, calibrations, and validations in each case are considered together, and the evaluation presented in Table 7-1 . These include: data needed, required amount of direct observations, suitable resolution, accuracy (under different missing types and input data types), analysis interval, computation costs, and validation status. Their definitions and descriptions are as follows:

Data needed: The data source is the first thing to consider. This criterion considers the types of sources needed for input. A circle (O) shows that the specific source is required by an approach or a method, and an x (×) means that it is not necessary.

Required amount of direct observations: In addition to the types of sources, the quantity is also a problem. Integration 2 needs a large amount of data to perform slightly more accurately than Integration 1, while it performs worse when there is less data. This goes to the discussion of the criterion for required direct observations: in the majority of the methods, they all need at least some direct observations to make estimations directly (such as Approaches 1 and 2), to form relationships (such as approach 3), and to calibrate the parameters (such as in Approach 4). As regression tools require much larger data than others, Approach 4 MLR and Integrated Method 2 for advanced MLR both call for larger direct flow observations.

Suitable resolution: The different methods involve different degrees of sensitivity and abilities to deal with data with different resolutions. Some approaches can only deal with the problem with lower resolution. For example, this thesis applies Approach 3 at 30 mins resolution due to the characteristics of FCD records.

Accuracy: In the previous analysis, the methods were mainly judged by error indicators. Standards are set here: The performance showing accuracy is defined by relative error MAPE (excellent: MAPE<=10%, good: 10 %< MAPE<=25%, Fair: 25 %< MAPE<=35%, Poor: MAPE> 35%). The accuracy is expressed under two concerns: Missing types and input data types.

- Missing types: This criterion shows whether the flows are missing during a long or short period, and during daytime or night. Some methods have distinguished performances in relation to this concern. For example, A4 MLR performs much better during the daytime than at night, and A1.2 can only be used during the night when the adaptive control has been turned on.

- Input data types: The original data inputs represent the original observations, and the processed data inputs represent the observations that have been smoothed. The difference in methods is revealed by this factor, according to underlying mechanisms: Approach 1 relies on the stability of flows in the time

dimension, while using the processed data after smoothing decreases the number of errors in the time dimension. This is not the case for approach 2. Hence, the accuracy of approach 1 is greatly improved after applying processed data.

Parameter analysis interval: The methods define their rules within a certain time range. For instance, the historical pattern can make an imputation for each single value with the same rules for each missing data point, but MLR trains its parameters every 4 h or 8 h, and applies these rules to make estimations. An iteration applies to each missing data point various rules, with different weights from observations of time and spatial dimensions. Thus, theoretically, the interval can range from 5 min. (the minimum observation interval) to 24h (a whole day).

Computation cost: Since a good method should be efficient, the computation cost is another significant criterion. For instance, although integrated method 1 shows good performance, performing the iteration takes a lot of time. The author measures the computation cost by the duration of calculation using a set of missing flows (from 0:00 to 24:00, one day, in one lane, with all experiments carried out using Matlab version R2014B). The standards are set as: Low: less than 1 second; medium: more than 1 second and less than 20 seconds; high: more than 20 seconds.

Validation status: Individual approaches 1, 2, and 4 are widely applied to cases on different lanes and days (see Appendices 1, 2, and 4). They can be seen as being validated. Besides, the two integrated methods, Iteration and Advanced MLR are validated in section 6.8.

Table 7-1. Comments on all of the approaches and integrated methods.

Approach/method	Sub event	Approach 1(1.1+1.2) Historical pattern	Approach 2 Lane distribution in timing plan	Approach 3 Fusing with FCD	Approach 4 Multiple linear regression	Integrated method 1- Iteration	Integrated method 2 –advanced regression
Data needed	Direction flow observation	O	O	O	O	O	O
	Timing plan	X(1.1) O(1.2)	O	X	X	O	O
	FCD	X	X	O	X	X	X
Required amount of direct observations		Low	Low	low	Large	low	Large
Suitable resolution		5/15/30minute interval	5/15/30minute interval	30minute interval	5/15/30minute interval	5/15/30minute interval	5/15/30minute interval
Missing types	Long-term	Good(1.1)	Good	/	Fair	Good	Good
	Short-term	Fair(1.1) Poor(1.2)	Good	/	Good(day time) Poor(night)	Excellent	Excellent(day time) Poor(night)
Input data types	Original data (raw data)	Fair	Fair	Poor(3.1) Fair(3.2)	Fair	Good	Good
	Processed data(smoothed data)	Good	Good	/	Good	Excellent	Excellent
Parameter analysis interval		/	/	<24 hours	> 4 hours < 24 hours	5 minute to 24 hours	> 4 hours < 24 hours
Computation costs		Low	Low	Medium	Low	High	Low
Validation status		Validated	Validated	/	Validated	Validated	Validated

Overall conclusion

This thesis starts from observations of traffic flow data from SCATS, and finds that the flows are widely missing from the system. This problem becomes the research question, "What are the best ways to estimate the traffic flow values missing at a junction?" In answering this question, many individual approaches are tried, from multiple angles, by applying available data sources: directly observed flows, timing plan, and FCD. The primary algorithms for the approaches are tested; they are then combined with other current algorithms, according to the results of analysis of the flow data, to develop new approaches. At the same time, the data fusion concept is used. Other sources, the timing plan, and FCD are linked with flows to determine pertinent relationships. After the processing of data and experiments from multiple scenarios involving certain key influential factors, historical (A1.1) and spatial (A2) relationships are shown to satisfy the assumptions. These are utilized in the estimation of missing flows. MLR (A4) has also been analysed; some recommendations are made according to its input and analysis interval. The approach A1.2 of using the green/flow ratio cannot be proven due to the limitations of the data sources. The approach of fusing FCD (A3.1) using speed-flow relations holds large errors, which still need to be improved. Inferences are then drawn in conclusion about reasons for failure, and recommendations are made for further research. The relationships of FCD counts-flow (A3.2) seem to be positive (they give better results, but still with large errors). Two integrations are demonstrated to show the possible combinations of individual approaches. These methods have better performances than the individual ones. They are also validated at another junction, and are thus confirmed as the best choices for the flow estimation at an urban junction. Finally, a grand comparison among all of the methods is conducted, showing the details of implementation of each approach or integrated method.

Recommendations for the application in practice

The series of methods presented in this thesis are suitable in multiple situations for all of the urban areas. When traffic flows are missing from a system, a specific choice of approaches should be considered, with the major concerns being:

- *Sources available: availability, and whether the data are direct observations or are drawn from other sources.*
- *Input/output requirements: Missing type (are data missing for a long- or short-term?); resolution (5/15/30 minutes); original or processed data.*
- *Other factors: analytical tools; computation costs, etc.*

For urban traffic systems like SCATS, considering the missing flows on a large scale, the calculation of the missing values has to be fast and efficient. In these cases, the approaches and methods presented in this thesis are highly recommended. The specific steps are as follows:

- *First, look at the available sources, including both direct flow observations and other data sources.*
- *Secondly, test the scales and types of missing data: To what extent for the missing flows are there missing values in the short or long-term, and do they appear during the daytime or at nights.*
- *Thirdly, positioning the target for the estimation: What are the concerned inputs (such as resolution/smoothed or original)? What are the required outputs (such as resolution)? How fast does it need to be (according to the scale and type of missing values)?*
- *Finally, make an initial choice of methods according to Table 7-1. Implement them to obtain primary results and comparisons, using integrated methods where possible. (If time is unlimited, the iteration method is recommended as providing more reliable results. If the inputs are enough, an advanced multiple linear regression is recommended.)*

Suggestions for future research

In this thesis, a series of methods are proposed for the estimation of missing flows at the urban junctions. However, it leaves some areas still to be investigated.

First, regarding MLR, two important factors are discussed in this thesis: The analysis interval for calibration and the relevant inputs. There is a question of choice between these two factors: a large analysis interval requires many inputs, wide data inputs will cause a lack of relevance, and with fewer inputs there are no reliable parameters. Hence, the increase of one will cause a negative impact from another. Therefore, there is a need for a proper analysis interval with a proper set of inputs. This will be done in further research.

Secondly, the case studies focus on one signal-controlled junction. However, the methods may also suitable for multiple adjacent junctions. For example, in approach 3, with data fusion using FCD and loop flow, the trajectories of FCD between two junctions can link two junctions together.

Besides, considering FCD speed and loop flow, when the flows are higher, the speeds become lower. This approach does not show much accuracy in the final estimation results, and a more precise means of expressing speed and flow will probably lead to more reliable results. The FCD counts and the total flow show a direct ratio, and the estimated result for missing values is around 30% on the stream level. This approach can be improved upon by looking into the dynamic penetration rate.

Finally, the methods in this thesis work are mainly tested for static traffic patterns. Whether they work for dynamic traffic patterns remains to be verified.

Acknowledgement

The thesis work is conducted in with TNO (Netherlands Organisation for Applied Scientific). Data used in the thesis come from Changsha Traffic Police.

Bibliography

Albright, D. (1991). Traffic Volume Summary Statistics. Transportation Research Record, (1305).

Antoniou, C., Balakrishna, R., & Koutsopoulos, H. N. (2011). A synthesis of emerging data collection technologies and their impact on traffic management applications. European Transport Research Review, 3(3), 139-148.

Banks, J. (2006). Part 1: traffic flow theory and characteristics: effect of time gaps and lane flow distributions on freeway bottleneck capacity.Transportation Research Record: Journal of the Transportation Research Board, (1965), 3-11.

Boyles, S. (2011). Comparison of Interpolation Methods for Missing Traffic Volume Data. In Transportation Research Board 90th Annual Meeting (No. 11-3757).

Cathey, F. W., & Dailey, D. J. (2003). Estimating corridor travel time by using transit vehicles as probes. Transportation Research Record: Journal of the Transportation Research Board, 1855(1), 60-65.

Chen, C., Kwon, J., Rice, J., Skabardonis, A., & Varaiya, P. (2003). Detecting errors and imputing missing data for single-loop surveillance systems.Transportation Research Record: Journal of the Transportation Research Board, 1855(1), 160-167.

Chen, D., Chen, L., & Liu, J. (2013). Road link traffic speed pattern mining in probe vehicle data via soft computing techniques. Applied Soft Computing,13(9), 3894-3902.

Chen, P., Nakamura, H., & Asano, M. (2012). Lane utilization analysis of shared left-turn lane based on saturation flow rate modeling. Procedia-Social and Behavioral Sciences, 43, 178-191.

Courage, K., Stephens, B., Gan, A., & Willis, M. (2002). Triple left-turn lanes at signalized intersections (No. Final Report).

Daamen, W., Buisson, C., & Hoogendoorn, S. P. (Eds.). (2014). Traffic Simulation and Data: Validation Methods and Applications. CRC Press.

Dailey, D. J., Harn, P., & Lin, P. J. (1996). ITS data fusion (No. WA-RD 410.1). Washington State Department of Transportation.

Deng, W., Lei, H., & Zhou, X. (2013). Traffic state estimation and uncertainty quantification based on heterogeneous data sources: A three detector approach. Transportation Research Part B: Methodological, 57, 132-157.

Fazio, J., Hoque, M., & Tiwari, G. (1999). Fatalities of heterogeneous street traffic. Transportation Research Record: Journal of the Transportation Research Board, (1695), 55-60.

Gerlough, D. L., & Huber, M. J. (1975). Traffic flow theory.

Hall, D. L., & Llinas, J. (1997). An introduction to multisensor data fusion.Proceedings of the IEEE, 85(1), 6-23.

Hamilton, J. D. (1994). Time series analysis (Vol. 2). Princeton: Princeton university press.

Huang, E., Antoniou, C., Wen, Y., Ben-Akiva, M., Lopes, J., & Bento, J. (2009, October). Real-time multi-sensor multi-source network data fusion using dynamic traffic assignment models. In Intelligent Transportation Systems, 2009. ITSC'09. 12th International IEEE Conference on (pp. 1-6). IEEE.

Jie, L., Van Zuylen, H., Chunhua, L., & Shoufeng, L. (2011). Monitoring travel times in an urban network using video, GPS and Bluetooth. Procedia-Social and Behavioral Sciences, 20, 630-637.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. Journal of Fluids Engineering, 82(1), 35-45.

Kergaye, C., Stevanovic, A., & Martin, P. (2009). Comparison of Before-After Versus Off-On Adaptive Traffic Control Evaluations: Park City, Utah, Case Study. Transportation Research Record: Journal of the Transportation Research Board, (2128), 192-201.

Klein, L. A., Mills, M. K., & Gibson, D. R. (2006). Traffic Detector Handbook: -Volume II (No. FHWA-HRT-06-139).

Kumar, S., Vanajakshi, L., & Subramanian, S. (2011). Location-Based Data for Estimated Traffic on Urban Arterial in Heterogeneous Traffic Conditions.Transportation Research Record: Journal of the Transportation Research Board, (2239), 16-22.

Kwon, T. M. (2004). TMC Traffic Data Automation for MnDOT's Traffic Monitoring Program.

Li, S., Zhu, K., Van Gelder, B., Nagle, J., & Tuttle, C. (2002). Reconsideration of sample size requirements for field traffic data collection with global positioning system devices. Transportation Research Record: Journal of the Transportation Research Board, (1804), 17-22.

Little, R. J., & Rubin, D. B. (2014). Statistical analysis with missing data. John Wiley & Sons.

Lowrie, P. R. (1990). Scats, sydney co-ordinated adaptive traffic system: A traffic responsive method of controlling urban traffic.

Lu, S., Li, J., & van Zuylen, H. (2012). The Evaluation of Traffic Control in Changsha City. Procedia-Social and Behavioral Sciences, 43, 216-225.

McCord, M., Mishalani, R., Goel, P., & Strohl, B. (2010). Iterative proportional fitting procedure to determine bus route passenger origin-destination flows. Transportation Research Record: Journal of the Transportation Research Board, (2145), 59-65.

Mendenhall, W., & Sincich, T. (1991). Statistics for Engineering and the Sciences. Dellen Pub. Co.. Collier Macmillan Canada. Maxwell Macmillan International.

Nakatsuji, T., Nakano, K., Nanthawichit, C., & Suzuki, H. (2004). Estimation of turning movements at intersections: Joint trip distribution and traffic assignment program combined with a genetic algorithm. Transportation Research Record: Journal of the Transportation Research Board, (1882), 53-60.

Nantes, A., Ngoduy, D., Bhaskar, A., Miska, M., & Chung, E. (2015). Real-time traffic state estimation in urban corridors from heterogeneous data. Transportation Research Part C: Emerging Technologies.

Nguyen, L. N., & Scherer, W. T. (2003). Imputation techniques to account for missing data in support of intelligent transportation systems applications (No. UVACTS-13-0-78,).

Ni, D., Leonard, J. D., Guin, A., & Feng, C. (2005). Multiple imputation scheme for overcoming the missing values and variability issues in ITS data. Journal of transportation engineering, 131(12), 931-938.

Ou, Q. (2011). Fusing Heterogeneous Traffic Data: Parsimonious Approaches using Data-Data Consistency. TU Delft, Delft University of Technology.

Petri, G. (2012). Information and dynamics in urban traffic networks (Doctoral dissertation, Imperial College London).

Qu, L., Li, L., Zhang, Y., & Hu, J. (2009). PPCA-based missing data imputation for traffic flow volume: a systematical approach. Intelligent Transportation Systems, IEEE Transactions on, 10(3), 512-522.

Remias, S., Hainen, A., Day, C., Brennan, T., Li, H., Rivera-Hernandez, E., ... & Bullock, D. (2013). Performance characterization of arterial traffic flow with probe

vehicle data. Transportation Research Record: Journal of the Transportation Research Board, (2380), 10-21.

Romana, M. (1999). Passing activity on two-lane highways in Spain.Transportation Research Record: Journal of the Transportation Research Board, (1678), 90-95.

Seo, T., Kusakabe, T., & Asakura, Y. (2015). Estimation of flow and density using probe vehicles with spacing measurement equipment. Transportation Research Part C: Emerging Technologies, 53, 134-150.

Shao, C. Q., Rong, J., & Liu, X. M. (2011). Study on the saturation flow rate and its influence factors at signalized intersections in China. Procedia-Social and Behavioral Sciences, 16, 504-514.

Smith, B., Scherer, W., & Hauser, T. (2001). Data-mining tools for the support of signal-timing plan development. Transportation Research Record: Journal of the Transportation Research Board, (1768), 141-147.

Stevanovic, A., Kergaye, C., & Stevanovic, J. (2012). Long-Term Benefits of Adaptive Traffic Control Under Varying Traffic Flows During Weekday Peak Hours. Transportation Research Record: Journal of the Transportation Research Board, (2311), 99-107.

Tang, J., Wang, Y., Zhang, S., Wang, H., & Liu, F. (2015). On Missing Traffic Data Imputation Based on Fuzzy C-means Method by Considering Spatial Temporal Correlation. In Transportation Research Board 94th Annual Meeting (No. 15-1334).

Tang, J., Zhang, G., Wang, Y., Wang, H., & Liu, F. (2015). A hybrid approach to integrate fuzzy C-means based imputation method with genetic algorithm for missing traffic volume data estimation. Transportation Research Part C: Emerging Technologies, 51, 29-40.

Taylor, M. A., Young, W., & Bonsall, P. W. (1996). Understanding traffic systems: data, analysis and presentation.

Treiber, M. Helbing, D., 2002. Reconstructing the spatio-temporal traffic dynamics from stationary detector data. Cooper@tive Tr@nsport@tion Dyn@mics 1, 3.1-3.24.

Turner, S., Albert, L., Gajewski, B., & Eisele, W. (2000). Archived intelligent transportation system data quality: Preliminary analyses of San Antonio TransGuide data. Transportation Research Record: Journal of the Transportation Research Board, (1719), 77-84.

Van Lint, J.W.C. Hoogendoorn, S.P., 2009. A robust and efficient method for fusing heterogeneous data from traffic sensors on freeways. Computer-Aided Civil and Infrastructure Engineering 25 (8), 596-612.

Van Zuylen, H. J., Zheng, F., & Chen, Y. (2010). An Investigation of Urban Link Travel Time Estimation Based on Field Sparse Probe Vehicle Data. In Transportation Research Board 89th Annual Meeting (No. 10-2522).

Varshney, P. K. (1997). Multisensor data fusion. Electronics & Communication Engineering Journal, 9(6), 245-253.

Wall, Z. R., & Dailey, D. J. (2003). Algorithm for detecting and correcting errors in archived traffic data. Transportation Research Record: Journal of the Transportation Research Board, 1855(1), 183-190.

Walpole, R. E., Myers, R. H., Myers, S. L., & Ye, K. (1993). Probability and statistics for engineers and scientists (Vol. 5). New York: Macmillan.

Wang, X., & Kockelman, K. (2009). Forecasting network data: Spatial interpolation of traffic counts from texas data. Transportation Research Record: Journal of the Transportation Research Board, (2105), 100-108.

Wang, Y., van Schuppen, J. H., & Vrancken, J. (2012). On-line distributed prediction of traffic flow in a large-scale traffic network. Proceedings of ITS World.

Wang, Y., van Schuppen, J. H., & Vrancken, J. (2014). Prediction of traffic flow at the boundary of a motorway network. Intelligent Transportation Systems, IEEE Transactions on, 15(1), 214-227.

Xiao, X., Chen, Y., & Yuan, Y. (2015). Estimation of Missing Flow at Junctions Using Control Plan and Floating Car Data. Transportation Research Procedia, 10, 113-123.

Xu, G., Huang, X., & Pant, P. (1999). Method for estimating capacity reduction in high-occupancy-vehicle lane ingress and egress sections. Transportation Research Record: Journal of the Transportation Research Board, (1678), 107-115.

Yuan, Y., Wilson, R. E., Van Lint, H., & Hoogendoorn, S. (2012). Estimation of Multiclass and Multilane Counts from Aggregate Loop Detector Data.Transportation Research Record: Journal of the Transportation Research Board, 2308(1), 120-127.

Zhang, X., Nihan, N., & Wang, Y. (2005). Improved dual-loop detection system for collecting real-time truck data. Transportation Research Record: Journal of the Transportation Research Board, (1917), 108-115.

Zhang, X., Wang, Y., Nihan, N., & Hallenbeck, M. (2003). Development of a system for collecting loop-detector event data for individual vehicles. Transportation Research Record: Journal of the Transportation Research Board, (1855), 168-175.

Zhong, M., Lingras, P. J., & Sharma, S. C. (2002). Applying short-term traffic prediction models for updating missing values of traffic counts.submitted to the Journal of Transportation Engineering, ASCE.

Appendix

Appendix 1 Historical pattern

Table 0-1 Estimation results using historical flow pattern for the second week on all lanes at junction 31616. Indicator MAPE, duration: the whole day (24 h), resolution 5 min. data input: original (raw) data

Lane#	Stream	Turning	MON	TUE	WED	THU	FRI	SAT	SUN	AVG Weekday	AVG Weekends	AVG all days
lane 1		Left	33.36%	35.54%	37.38%	34.90%	41.17%	31.24%	38.56%	36.47%	34.90%	36.02%
lane 2		Left	35.35%	46.09%	38.30%	37.90%	36.25%	31.39%	34.59%	38.78%	32.99%	37.12%
lane 3	West	Throughput	55.38%	43.28%	41.54%	36.80%	36.06%	32.35%	36.54%	42.61%	34.45%	40.28%
lane 4	stream	Throughput	43.52%	39.15%	41.14%	36.28%	31.42%	31.43%	33.98%	38.30%	32.71%	36.70%
lane 5		Throughput	37.43%	34.58%	35.85%	38.06%	35.09%	31.35%	36.22%	36.20%	33.79%	35.51%
lane 6		Right	82.65%	57.93%	77.41%	79.16%	68.87%	58.60%	67.45%	73.20%	63.03%	70.30%
lane 7		Left	29.91%	33.69%	43.85%	30.51%	27.13%	25.00%	29.70%	33.02%	27.35%	31.40%
lane 8		Left	29.85%	38.15%	31.03%	31.62%	36.22%	23.79%	28.60%	33.37%	26.20%	31.32%
lane 9	South	Throughput	32.07%	45.97%	41.35%	38.06%	27.88%	26.95%	35.19%	37.07%	31.07%	35.35%
lane 10	stream	Throughput	24.20%	31.17%	24.14%	27.75%	22.98%	20.54%	25.27%	26.05%	22.91%	25.15%
lane 11		Throughput	27.10%	24.49%	24.56%	26.25%	36.65%	19.27%	25.25%	27.81%	22.26%	26.22%
lane 12		Right	52.84%	45.42%	46.20%	48.46%	50.46%	37.99%	37.02%	48.68%	37.51%	45.48%
lane 13		Left	58.91%	40.94%	44.27%	41.03%	48.76%	44.15%	40.71%	46.78%	42.43%	45.54%
lane 14		Left	58.25%	43.38%	40.50%	37.64%	44.64%	38.41%	39.33%	44.88%	38.87%	43.16%
lane 15	East	Throughput	51.75%	35.54%	38.05%	34.60%	49.62%	37.75%	43.81%	41.91%	40.78%	41.59%
lane 16	stream	Throughput	46.57%	39.96%	42.93%	47.02%	45.52%	41.31%	44.53%	44.40%	42.92%	43.98%
lane 17		Throughput	34.04%	33.85%	34.69%	30.58%	34.74%	30.03%	41.69%	33.58%	35.86%	34.23%
lane 18		Right	75.01%	58.74%	58.79%	53.61%	59.62%	51.93%	41.92%	61.15%	46.93%	57.09%
lane 19		Left	48.73%	41.26%	42.01%	38.36%	51.88%	43.24%	51.50%	44.45%	47.37%	45.28%
lane 20		Left	44.58%	55.26%	69.19%	59.51%	44.25%	40.59%	37.13%	54.56%	38.86%	50.07%
lane 21	North	Throughput	250.29%	34.77%	34.33%	25.11%	28.52%	26.21%	53.84%	74.60%	40.03%	64.72%
lane 22	stream	Throughput	41.94%	45.24%	54.89%	44.72%	42.25%	47.09%	56.38%	45.81%	51.74%	47.50%
lane 23		Throughput	36.70%	26.65%	27.96%	26.09%	51.75%	37.92%	53.40%	33.83%	45.66%	37.21%
lane 24		Right	26.93%	33.96%	32.06%	22.09%	31.26%	28.73%	26.43%	29.26%	27.58%	28.78%
AVG Left			42.37%	41.79%	43.32%	38.93%	41.29%	34.73%	37.52%	41.54%	36.12%	39.99%
AVG Throughput			56.75%	36.22%	36.79%	34.28%	36.87%	31.85%	40.51%	40.18%	36.18%	39.04%
AVG Right			59.36%	49.01%	53.62%	50.83%	52.55%	44.31%	43.21%	53.07%	43.76%	50.41%
AVG West			47.95%	42.76%	45.27%	43.85%	41.48%	36.06%	41.22%	44.26%	38.64%	42.66%
AVG South			32.66%	36.48%	35.19%	33.78%	33.55%	25.59%	30.17%	34.33%	27.88%	32.49%
AVG East			54.09%	42.07%	43.21%	40.75%	47.15%	40.60%	42.00%	45.45%	41.30%	44.26%
AVG North			74.86%	39.52%	43.41%	35.98%	41.65%	37.30%	46.45%	47.08%	41.87%	45.60%
AVG Total lanes			52.39%	40.21%	41.77%	38.59%	40.96%	34.89%	39.96%	42.78%	37.42%	41.25%

Table 0-2 Estimation results using historical flow pattern for the second week on all lanes at junction 31616. Indicator MAPE, duration: the whole day (24 h), resolution 5 min. data input: processed (smoothed) data

Lane#	Stream	Turning	MON	TUE	WED	THU	FRI	SAT	SUN	AVG Weekday	AVG Weekends	AVG all days
lane 1		Left	16.32%	15.85%	19.13%	14.78%	19.22%	12.84%	18.57%	17.06%	15.71%	16.67%
lane 2		Left	21.96%	22.77%	22.05%	16.50%	20.58%	16.28%	16.00%	20.77%	16.14%	19.45%
lane 3	West	Throughput	40.19%	21.87%	23.47%	21.75%	21.70%	19.80%	22.77%	25.80%	21.29%	24.51%
lane 4	stream	Throughput	28.93%	21.34%	23.28%	18.92%	16.64%	15.36%	17.45%	21.82%	16.41%	20.27%
lane 5		Throughput	22.27%	16.68%	16.96%	20.33%	17.74%	15.86%	18.60%	18.80%	17.23%	18.35%
lane 6		Right	52.13%	46.15%	54.40%	45.41%	40.67%	34.65%	37.82%	47.75%	36.24%	44.46%
lane 7		Left	18.70%	16.99%	17.13%	12.85%	10.86%	13.20%	16.03%	15.31%	14.62%	15.11%
lane 8		Left	18.50%	19.86%	15.87%	14.77%	21.62%	11.12%	14.75%	18.12%	12.94%	16.64%
lane 9	South	Throughput	22.95%	31.75%	29.50%	23.30%	16.29%	11.87%	15.82%	24.76%	13.85%	21.64%
lane 10	stream	Throughput	15.35%	13.78%	13.23%	13.19%	10.48%	9.78%	11.38%	13.21%	10.58%	12.46%
lane 11		Throughput	13.81%	12.06%	11.72%	10.20%	23.78%	9.91%	11.82%	14.31%	10.87%	13.33%
lane 12		Right	35.08%	30.15%	28.11%	24.49%	30.31%	24.94%	25.06%	29.63%	25.00%	28.31%
lane 13		Left	41.05%	28.11%	26.17%	24.80%	30.54%	24.64%	27.54%	30.13%	26.09%	28.98%
lane 14		Left	43.46%	29.00%	23.03%	22.95%	27.41%	22.16%	24.21%	29.17%	23.19%	27.46%
lane 15	East	Throughput	39.71%	28.24%	27.60%	23.16%	46.85%	26.98%	28.02%	33.11%	27.50%	31.51%
lane 16	stream	Throughput	27.16%	24.04%	25.74%	24.64%	25.46%	23.23%	22.87%	25.41%	23.05%	24.73%
lane 17		Throughput	17.57%	18.33%	18.43%	16.76%	16.03%	18.50%	21.22%	17.42%	19.86%	18.12%
lane 18		Right	53.60%	37.71%	40.92%	33.16%	38.54%	33.04%	31.87%	40.79%	32.46%	38.41%
lane 19		Left	36.59%	22.87%	26.79%	25.76%	34.93%	34.40%	35.51%	29.39%	34.96%	30.98%
lane 20		Left	28.97%	37.78%	35.91%	32.87%	28.34%	22.46%	23.75%	32.77%	23.11%	30.01%
lane 21	North	Throughput	229.00%	21.26%	19.55%	13.83%	16.16%	20.91%	23.83%	59.96%	22.37%	49.22%
lane 22	stream	Throughput	20.11%	27.03%	26.32%	22.16%	17.09%	24.02%	33.15%	22.54%	28.59%	24.27%
lane 23		Throughput	22.13%	12.75%	15.04%	13.61%	36.62%	16.69%	21.97%	20.03%	19.33%	19.83%
lane 24		Right	18.54%	19.72%	17.38%	12.47%	15.53%	16.65%	14.09%	16.73%	15.37%	16.34%
AVG Left			28.19%	24.15%	23.26%	20.66%	24.19%	19.64%	22.05%	24.09%	20.84%	23.16%
AVG Throughput			41.60%	20.76%	20.90%	18.49%	22.07%	17.74%	20.74%	24.76%	19.24%	23.19%
AVG Right			39.84%	33.43%	35.20%	28.88%	31.26%	27.32%	27.21%	33.72%	27.27%	31.88%
AVG West			30.30%	24.11%	26.55%	22.95%	22.76%	19.13%	21.87%	25.33%	20.50%	23.95%
AVG South			20.73%	20.77%	19.26%	16.47%	18.89%	13.47%	15.81%	19.22%	14.64%	17.91%
AVG East			37.09%	27.57%	26.98%	24.25%	30.81%	24.76%	25.96%	29.34%	25.36%	28.20%
AVG North			59.22%	23.57%	23.50%	20.12%	24.78%	22.52%	25.38%	30.24%	23.95%	28.44%
AVG Total lanes			36.84%	24.00%	24.07%	20.94%	24.31%	19.97%	22.25%	26.03%	21.11%	24.63%

Appendix 2 Lane spatial distribution

Table 0-3 Estimation results using lane spatial distribution for the second week on all lanes at junction 31616. Indicator MAPE, duration: the whole day (24 h), resolution 5 min. data input: original (raw) data

Lane#	Stream	Turning	MON	TUE	WED	THU	FRI	SAT	SUN	AVG	AVG	AVG
										Weekday	Weekends	all days
lane 1		Left	24.96%	35.39%	24.57%	27.39%	33.79%	28.42%	27.72%	29.22%	28.07%	28.89%
lane 2		Left	31.00%	42.18%	28.17%	31.54%	31.87%	28.79%	29.95%	32.95%	29.37%	31.93%
lane 3	West	Throughput	44.23%	48.90%	36.62%	41.55%	40.63%	36.54%	39.65%	42.39%	38.10%	41.16%
lane 4	stream	Throughput	31.22%	37.32%	30.02%	27.80%	27.82%	27.85%	25.72%	30.84%	26.79%	29.68%
lane 5		Throughput	27.10%	32.35%	28.62%	30.89%	28.62%	28.38%	33.56%	29.52%	30.97%	29.93%
lane 6		Right	62.12%	57.82%	66.18%	63.92%	61.29%	53.97%	63.87%	62.27%	58.92%	61.31%
lane 7		Left	22.64%	24.54%	34.57%	23.80%	22.45%	24.21%	23.94%	25.60%	24.08%	25.16%
lane 8		Left	29.55%	30.73%	26.94%	30.08%	36.15%	27.44%	25.25%	30.69%	26.35%	29.45%
lane 9	South	Throughput	24.02%	38.06%	37.40%	33.53%	27.10%	33.04%	31.70%	32.02%	32.37%	32.12%
lane 10	stream	Throughput	18.10%	19.50%	19.23%	18.14%	17.01%	18.09%	17.67%	18.40%	17.88%	18.25%
lane 11		Throughput	20.00%	21.48%	22.12%	21.53%	21.57%	19.34%	17.03%	21.34%	18.19%	20.44%
lane 12		Right	46.33%	48.97%	49.99%	45.22%	43.61%	42.64%	39.36%	46.82%	41.00%	45.16%
lane 13		Left	44.39%	41.24%	41.49%	39.84%	38.05%	38.19%	34.98%	41.00%	36.59%	39.74%
lane 14		Left	40.07%	41.30%	35.91%	39.25%	37.61%	34.78%	33.03%	38.83%	33.91%	37.42%
lane 15	East	Throughput	35.35%	36.57%	34.95%	30.54%	40.65%	33.78%	33.45%	35.61%	33.62%	35.04%
lane 16	stream	Throughput	42.95%	40.09%	41.27%	48.71%	38.94%	43.30%	41.04%	42.39%	42.17%	42.33%
lane 17		Throughput	33.05%	36.95%	34.10%	37.00%	36.32%	38.72%	45.23%	35.48%	41.98%	37.34%
lane 18		Right	61.62%	60.88%	59.23%	58.37%	64.69%	55.65%	53.65%	60.96%	54.65%	59.16%
lane 19		Left	43.99%	34.95%	38.31%	34.63%	56.19%	47.38%	50.03%	41.61%	48.71%	43.64%
lane 20		Left	57.28%	67.01%	75.11%	75.01%	52.89%	45.53%	45.82%	65.46%	45.68%	59.81%
lane 21	North	Throughput	259.87%	20.12%	23.05%	21.65%	24.90%	26.38%	62.31%	69.92%	44.35%	62.61%
lane 22	stream	Throughput	40.60%	35.41%	43.24%	38.58%	37.00%	43.25%	51.34%	38.97%	47.30%	41.35%
lane 23		Throughput	31.76%	20.41%	19.59%	21.00%	41.20%	35.62%	55.02%	26.79%	45.32%	32.09%
lane 24		Right	22.88%	33.75%	30.44%	23.25%	30.47%	30.39%	27.18%	28.16%	28.79%	28.34%
AVG Left			36.74%	39.67%	38.13%	37.69%	38.63%	34.34%	33.84%	38.17%	34.09%	37.01%
AVG Throughput			50.69%	32.26%	30.85%	30.91%	31.81%	32.02%	37.81%	35.31%	34.92%	35.19%
AVG Right			48.24%	50.36%	51.46%	47.69%	50.02%	45.66%	46.02%	49.55%	45.84%	48.49%
AVG West			36.77%	42.33%	35.70%	37.18%	37.34%	33.99%	36.75%	37.86%	35.37%	37.15%
AVG South			26.77%	30.55%	31.71%	28.72%	27.98%	27.46%	25.83%	29.15%	26.64%	28.43%
AVG East			42.91%	42.84%	41.16%	42.29%	42.71%	40.74%	40.23%	42.38%	40.48%	41.84%
AVG North			76.06%	35.28%	38.29%	35.69%	40.44%	38.09%	48.62%	45.15%	43.35%	44.64%
AVG Total lanes			45.63%	37.75%	36.71%	35.97%	37.12%	35.07%	37.85%	38.63%	36.46%	38.01%

Table 0-4 Estimation results using lane spatial distribution for the second week on all lanes at junction 31616. Indicator MAPE, duration: the whole day (24 h), resolution 5 min. data input: processed (smoothed) data

Lane#	Stream	Turning	MON	TUE	WED	THU	FRI	SAT	SUN	AVG Weekday	AVG Weekends	AVG all days
lane 1		Left	15.21%	26.16%	15.48%	16.70%	21.57%	17.65%	17.41%	19.02%	17.53%	18.60%
lane 2		Left	21.73%	28.07%	18.71%	20.53%	25.08%	19.64%	19.19%	22.82%	19.42%	21.85%
lane 3	West	Throughput	37.87%	35.19%	28.28%	34.54%	37.19%	34.06%	35.38%	34.61%	34.72%	34.64%
lane 4	stream	Throughput	20.37%	24.97%	20.38%	17.41%	19.08%	18.92%	18.15%	20.44%	18.54%	19.90%
lane 5		Throughput	20.25%	22.77%	17.43%	20.26%	21.10%	21.53%	23.38%	20.36%	22.46%	20.96%
lane 6		Right	39.43%	39.86%	40.86%	39.94%	36.60%	33.22%	41.77%	39.34%	37.50%	38.81%
lane 7		Left	13.68%	17.18%	14.62%	12.24%	12.48%	14.62%	15.84%	14.04%	15.23%	14.38%
lane 8		Left	16.92%	20.19%	16.37%	17.46%	27.95%	14.59%	15.03%	19.78%	14.81%	18.36%
lane 9	South	Throughput	17.55%	32.72%	31.14%	26.02%	20.47%	19.80%	17.09%	25.58%	18.45%	23.54%
lane 10	stream	Throughput	13.65%	12.89%	14.89%	13.18%	13.22%	13.81%	13.43%	13.57%	13.62%	13.58%
lane 11		Throughput	12.67%	19.50%	16.42%	14.18%	13.01%	14.81%	13.51%	15.16%	14.16%	14.87%
lane 12		Right	29.51%	32.67%	31.53%	23.06%	26.21%	29.84%	25.63%	28.60%	27.74%	28.35%
lane 13		Left	28.53%	29.96%	27.65%	26.17%	24.94%	26.88%	26.23%	27.45%	26.56%	27.19%
lane 14		Left	29.06%	31.19%	26.30%	28.05%	28.24%	25.63%	24.09%	28.57%	24.86%	27.51%
lane 15	East	Throughput	27.70%	30.61%	29.27%	25.25%	43.34%	28.71%	26.07%	31.23%	27.39%	30.14%
lane 16	stream	Throughput	29.94%	28.87%	32.05%	35.17%	28.94%	32.81%	29.78%	30.99%	31.30%	31.08%
lane 17		Throughput	27.87%	30.36%	28.27%	28.58%	28.76%	30.59%	35.96%	28.77%	33.28%	30.06%
lane 18		Right	46.26%	42.91%	43.87%	42.10%	45.70%	43.42%	38.88%	44.17%	41.15%	43.31%
lane 19		Left	38.50%	24.62%	28.83%	29.93%	46.32%	42.25%	39.94%	33.64%	41.10%	35.77%
lane 20		Left	44.52%	58.86%	54.21%	53.18%	46.49%	33.46%	35.79%	51.45%	34.63%	46.64%
lane 21	North	Throughput	264.75%	14.21%	17.28%	17.30%	18.90%	23.73%	36.30%	66.49%	30.02%	56.07%
lane 22	stream	Throughput	24.10%	25.27%	22.68%	22.24%	22.98%	27.34%	35.96%	23.45%	31.65%	25.80%
lane 23		Throughput	21.68%	13.94%	14.49%	14.02%	36.31%	16.96%	29.95%	20.09%	23.46%	21.05%
lane 24		Right	14.08%	18.90%	19.27%	15.90%	15.51%	18.73%	17.52%	16.73%	18.13%	17.13%
AVG Left			26.02%	29.53%	25.27%	25.53%	29.13%	24.34%	24.19%	27.10%	24.27%	26.29%
AVG Throughput			43.20%	24.28%	22.72%	22.35%	25.28%	23.59%	26.25%	27.56%	24.92%	26.81%
AVG Right			32.32%	33.59%	33.88%	30.25%	31.01%	31.30%	30.95%	32.21%	31.13%	31.90%
AVG West			25.81%	29.50%	23.52%	24.90%	26.77%	24.17%	25.88%	26.10%	25.03%	25.79%
AVG South			17.33%	22.53%	20.83%	17.69%	18.89%	17.91%	16.76%	19.45%	17.33%	18.85%
AVG East			31.56%	32.32%	31.24%	30.89%	33.32%	31.34%	30.17%	31.86%	30.75%	31.55%
AVG North			67.94%	25.97%	26.13%	25.43%	31.09%	27.08%	32.58%	35.31%	29.83%	33.74%
AVG Total lanes			35.66%	27.58%	25.43%	24.73%	27.52%	25.13%	26.35%	28.18%	25.74%	27.48%

Appendix 3 FCD-loop flow relation

Speed-flow relation in approach 3 (FCD-flow data fusion)

- ***Inbound area***

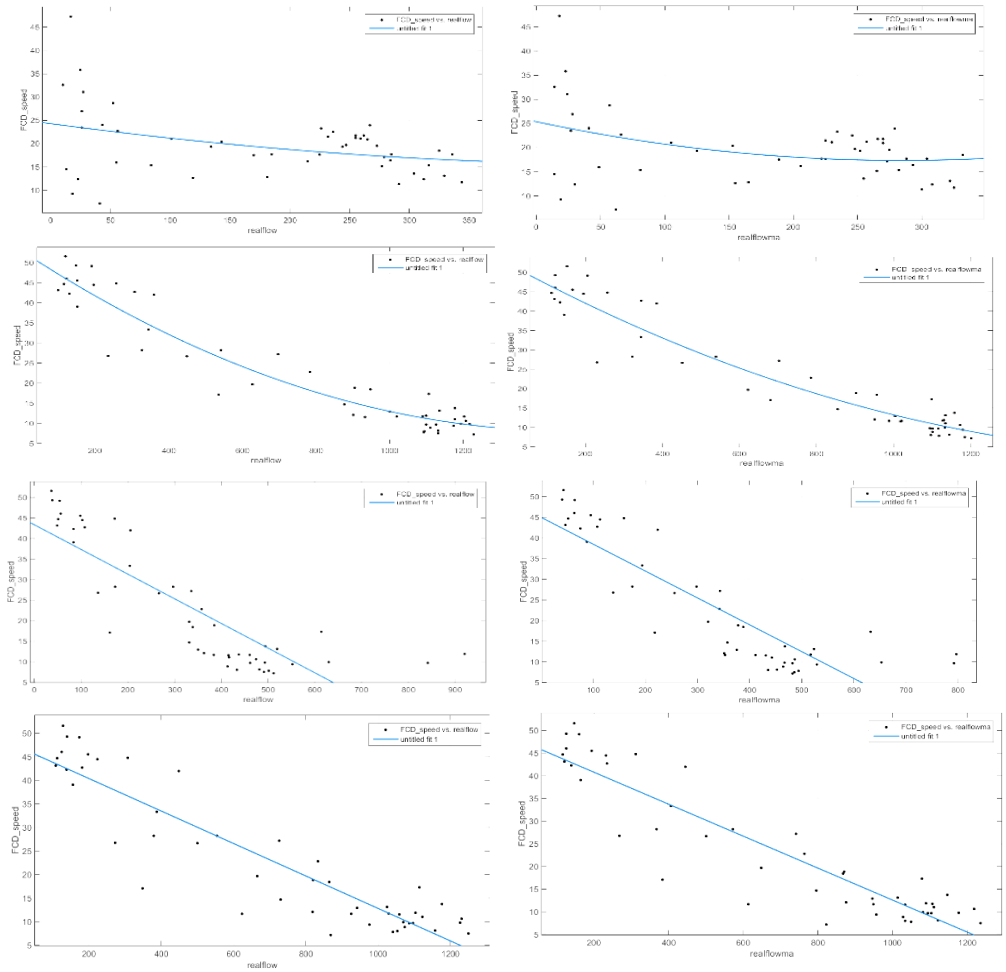


Figure 0-1 The FCD speed- loop flow relation formed from inbound from streams (Top-down – East, South, West, North)

Table 0-5 fitting parameters from inbound, junction 31616, 23rd April 2013

	East		**South**		**West**		**North**	
Data	Original	Smoothed	Original	Smoothed	Original	Smoothed	Original	Smoothed
Fitting	f(x) = p1*x^2 + p2*x + p3				f(x) = p1*x+ p2		f(x) = p1*x^2 + p2*x + p3	
p1	3.34e-05	9.91e-05	2.01e-05	1.46e-05	-0.05	-0.06	3.34e-05	9.91e-05
p2	-0.03	-0.05	-0.06	-0.05	43.32	44.99	-0.03	-0.05
p3	24.31	25.37	52.96	52.18			47.20	47.84
Goodness of fit								
R-square	0.14	0.15	0.91	0.91	0.70	0.75	0.84	0.84
RMSE	6.75	6.70	4.43	4.36	8.13	7.44	6.00	5.86

- ***Outbound area***

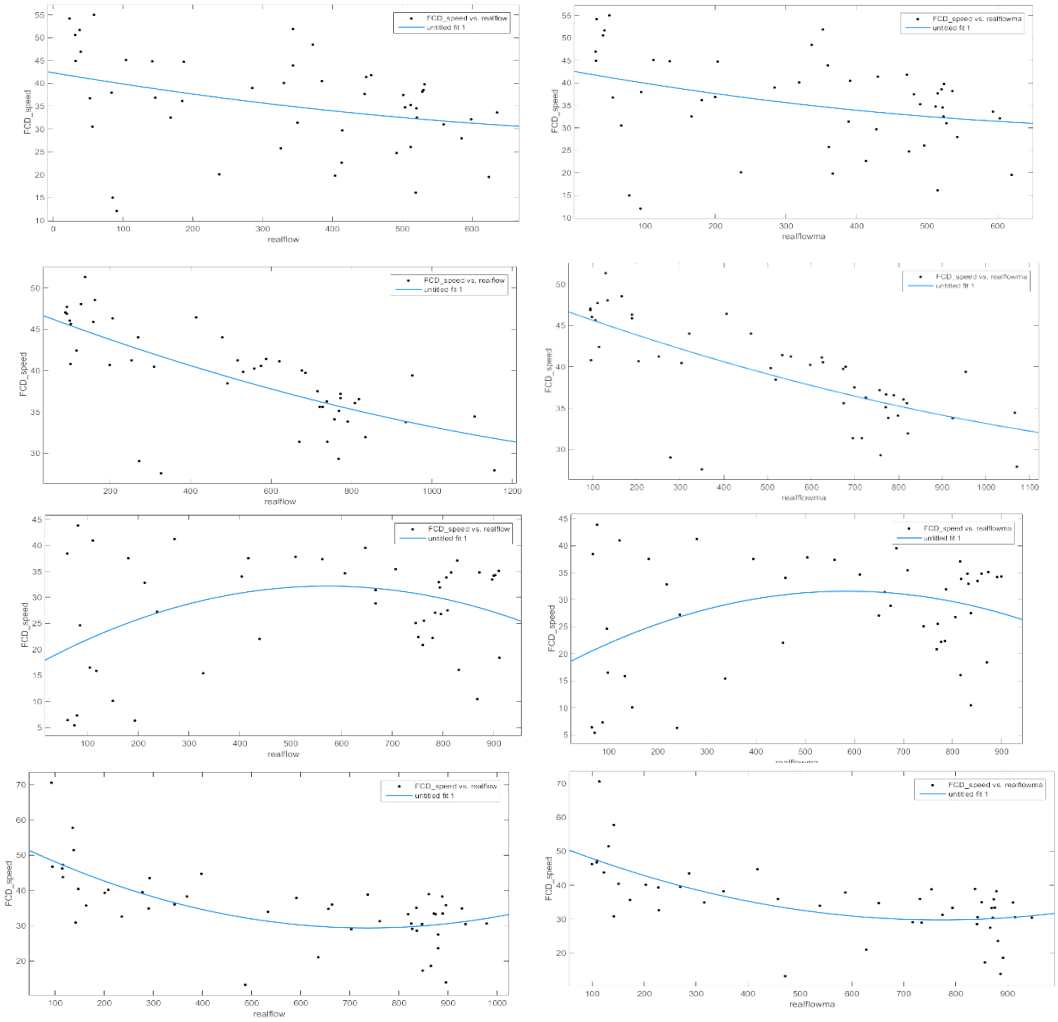


Figure 0-2 The FCD speed- loop flow relation formed from outbound from streams (Top-down – East, South, West, North)

Table 0-6 fitting parameters from outbound, junction 31616, 23rd April 2013

	East		**South**		**West**		**North**	
	Original data	Smoothed data	Original data	Smoothed data	Original data	Smoothed data	Original data	Smoothed data
Fitting curve	f(x) = p1*x^2 + p2*x + p3							
p1	1.22e-05	1.54e-05	4.33e-06	4.76e-06	-4.63e-05	-4.10e-05	4.63e-05	4.02e-05
p2	-0.02	-0.02	-0.02	-0.02	0.05	0.04	-0.07	-0.06
p3	42.32	42.55	47.27	47.51	17.05	17.55	54.48	53.67
Goodness of fit								
R-square	0.12	0.12	0.51	0.51	0.12	0.10	0.45	0.43
RMSE	10	10.02	4.22	4.21	10.18	10.27	7.84	7.97

Count-flow relation in approach 3 (FCD-flow data fusion)

- ***Inbound area***

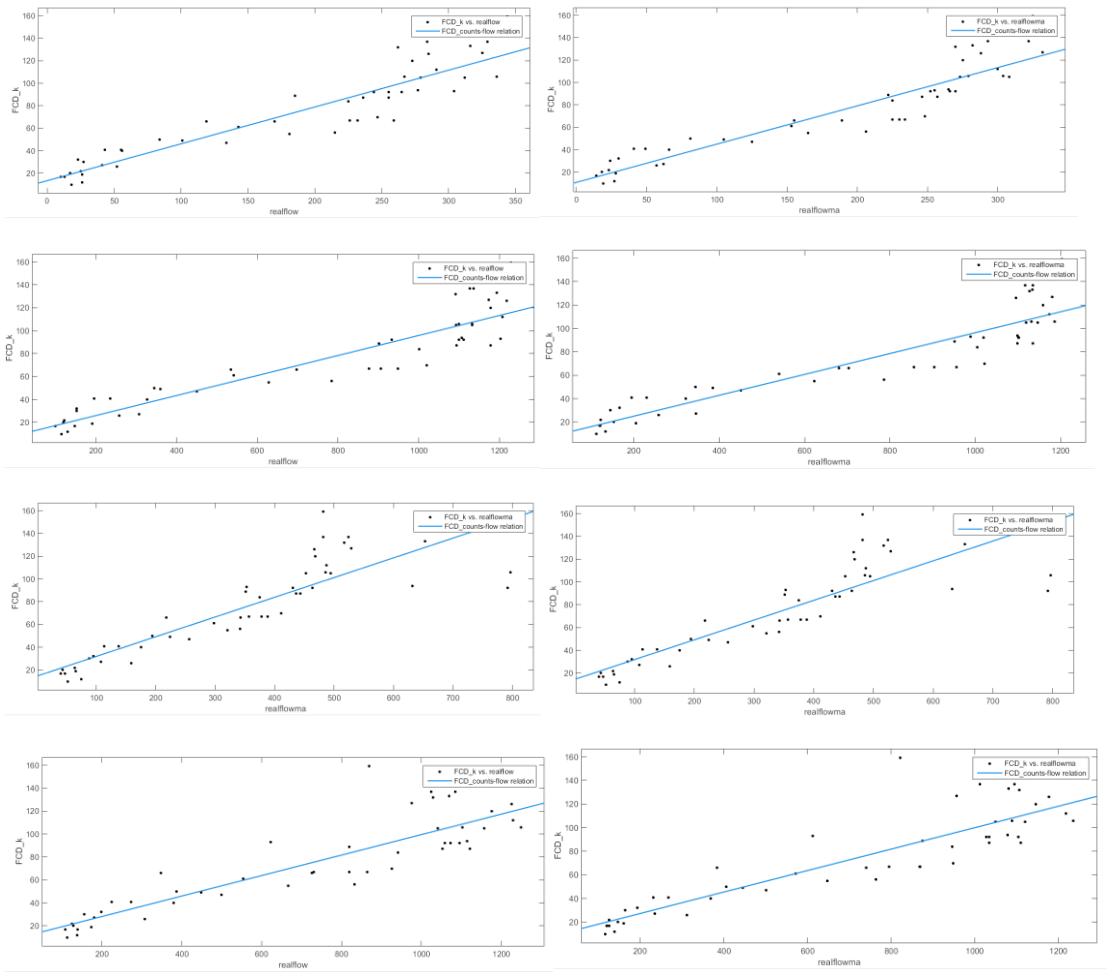


Figure 0-3 The FCD count- loop flow relation formed from inbound from streams (Top-down – East, South, West, North)

Table 0-7 fitting parameters from inbound, junction 31616, 23rd April 2013

	East		**South**		**West**		**North**	
	Original data	Smoothed data	Original data	Smoothed data	Original data	Smoothed data	Original data	Smoothed data
Fitting curve	f(x) = p1*x+ p2							
p1	0.32	0.34	0.09	0.09	0.16	0.17	0.08	0.09
p2	13.46	10.85	8.55	7.08	19.12	14.56	10.30	9.14
Goodness of fit								
R-square	0.86	0.87	0.85	0.85	0.70	0.74	0.77	0.78
RMSE	15.05	14.15	15.51	15.54	22.35	20.44	18.96	18.92

- ***Outbound area***

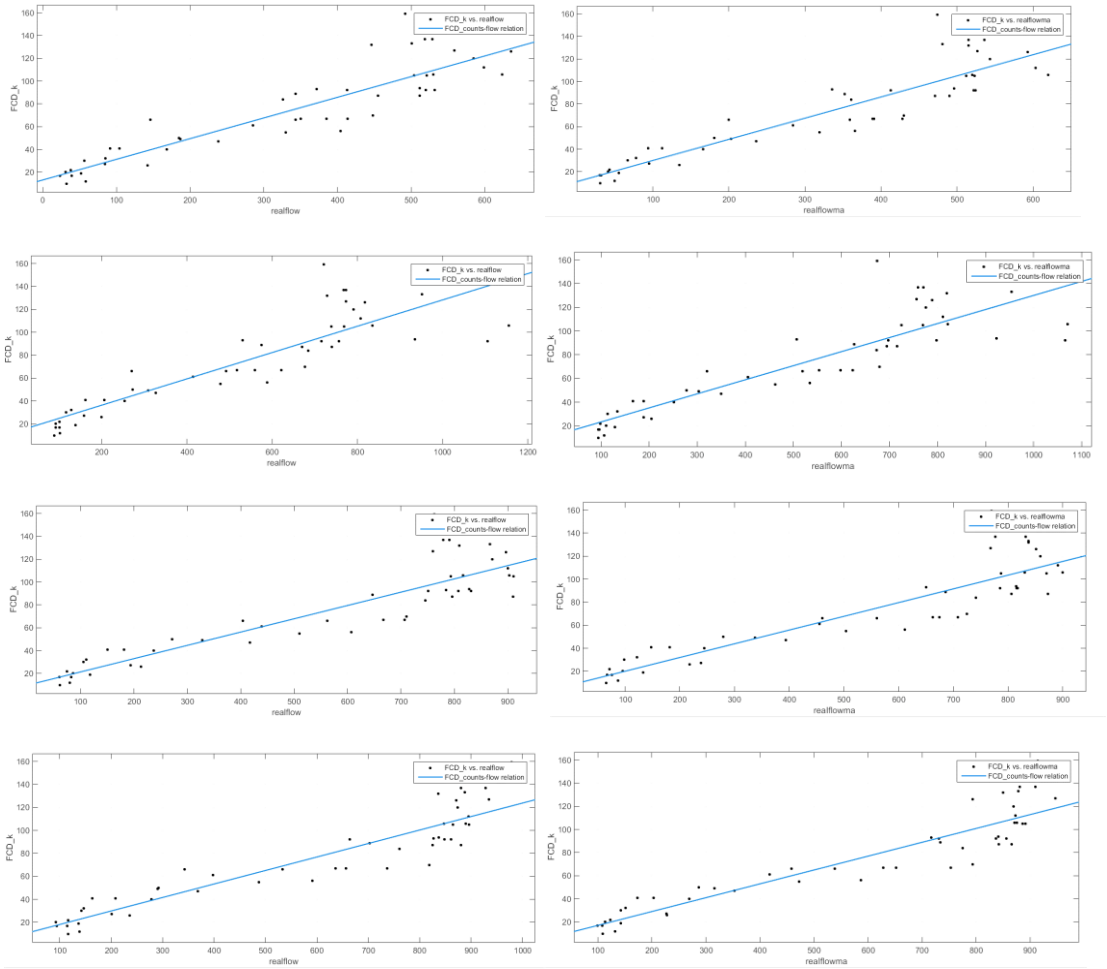


Figure 0-4 The FCD count- loop flow relation formed from outbound from direction streams (Top-down – East, South, West, North)

Table 0-8 fitting parameters from outbound, junction 31616, 23rd April 2013

	East		**South**		**West**		**North**	
	Original data	Smoothed data	Original data	Smoothed data	Original data	Smoothed data	Original data	Smoothed data
Fitting curve	f(x) = p1*x+ p2							
p1	0.19	0.19	0.11	0.12	0.11	0.12	0.11	0.12
p2	13.16	11.16	13.37	11.39	9.71	8.09	6.39	5.01
Goodness of fit								
R-square	0.82	0.83	0.76	0.78	0.81	0.82	0.86	0.86
RMSE	17.26	16.69	19.52	19.03	17.71	17.37	14.89	15.09

Appendix 4 MLR

Scenario 1: input range: lanes from the whole junction, 24h analysis interval

Table 0-9 Estimation results using MLR for the second week on all lanes at junction 31616. Indicator MAPE, duration: the whole day (24 h), resolution 5 min. analysis interval: 24 h, inputs category: all the lanes at a junction and a whole week, data input: original (raw) data

Lane#	Stream	Turning	MON	TUE	WED	THU	FRI	SAT	SUN	AVG Weekday	AVG Weekends	AVG all days
lane 1		Left	42.68%	35.65%	40.50%	32.39%	46.14%	33.27%	35.87%	39.47%	34.57%	38.07%
lane 2		**Left**	**45.79%**	**40.28%**	**39.98%**	**39.62%**	**34.74%**	**36.81%**	**38.33%**	**40.08%**	**37.57%**	**39.36%**
lane 3	West	Throughput	56.90%	46.24%	47.60%	48.71%	48.09%	46.27%	47.87%	49.51%	47.07%	48.81%
lane 4	stream	**Throughput**	**38.46%**	**44.39%**	**45.61%**	**41.25%**	**43.52%**	**41.92%**	**36.12%**	**42.65%**	**39.02%**	**41.61%**
lane 5		Throughput	43.61%	44.84%	49.94%	52.04%	44.52%	42.85%	42.48%	46.99%	42.66%	45.75%
lane 6		**Right**	**114.36%**	**108.86%**	**107.68%**	**104.96%**	**109.95%**	**87.80%**	**79.84%**	**109.16%**	**83.82%**	**101.92%**
lane 7		Left	27.22%	39.90%	43.35%	34.24%	29.18%	43.38%	31.63%	34.78%	37.50%	35.56%
lane 8		**Left**	**38.80%**	**37.66%**	**45.30%**	**33.33%**	**32.04%**	**41.09%**	**29.84%**	**37.42%**	**35.46%**	**36.86%**
lane 9	South	Throughput	32.44%	45.28%	41.77%	38.39%	36.01%	32.69%	32.14%	38.78%	32.42%	36.96%
lane 10	stream	**Throughput**	**21.61%**	**27.74%**	**26.35%**	**28.84%**	**23.60%**	**22.27%**	**22.15%**	**25.63%**	**22.21%**	**24.65%**
lane 11		Throughput	27.78%	25.59%	28.16%	26.52%	33.05%	19.63%	20.86%	28.22%	20.24%	25.94%
lane 12		**Right**	**94.86%**	**84.98%**	**95.30%**	**72.66%**	**75.32%**	**79.63%**	**62.90%**	**84.62%**	**71.26%**	**80.81%**
lane 13		Left	63.48%	67.01%	76.32%	63.17%	67.96%	61.10%	55.39%	67.59%	58.25%	64.92%
lane 14		**Left**	**63.20%**	**65.81%**	**59.24%**	**53.66%**	**55.00%**	**58.76%**	**62.64%**	**59.38%**	**60.70%**	**59.76%**
lane 15	East	Throughput	73.61%	61.98%	65.37%	42.46%	64.60%	72.48%	56.27%	61.60%	64.38%	62.40%
lane 16	stream	**Throughput**	**56.24%**	**63.08%**	**72.10%**	**56.30%**	**67.83%**	**62.07%**	**48.78%**	**63.11%**	**55.42%**	**60.91%**
lane 17		Throughput	37.83%	45.05%	44.27%	43.22%	50.93%	46.55%	46.63%	44.26%	46.59%	44.93%
lane 18		**Right**	**90.39%**	**102.58%**	**90.27%**	**92.58%**	**99.50%**	**81.96%**	**89.83%**	**95.06%**	**85.90%**	**92.45%**
lane 19		Left	87.78%	95.06%	71.52%	76.03%	74.29%	70.18%	66.83%	80.94%	68.51%	77.39%
lane 20		**Left**	**73.63%**	**96.81%**	**115.05%**	**89.04%**	**80.67%**	**72.41%**	**77.30%**	**91.04%**	**74.85%**	**86.41%**
lane 21	North	Throughput	240.30%	46.53%	31.16%	44.23%	26.69%	30.43%	51.96%	77.78%	41.20%	67.33%
lane 22	stream	**Throughput**	**55.97%**	**59.85%**	**59.55%**	**64.53%**	**64.33%**	**59.75%**	**60.94%**	**60.85%**	**60.35%**	**60.70%**
lane 23		Throughput	60.05%	47.77%	45.80%	33.29%	47.42%	37.75%	56.51%	46.87%	47.13%	46.94%
lane 24		**Right**	**47.48%**	**44.42%**	**39.93%**	**36.79%**	**42.77%**	**37.92%**	**33.22%**	**42.28%**	**35.57%**	**40.36%**
AVG Left			55.32%	59.77%	61.41%	52.68%	52.50%	52.12%	49.73%	56.34%	50.93%	54.79%
AVG Throughput			**62.07%**	**46.53%**	**46.47%**	**43.31%**	**45.88%**	**42.89%**	**43.56%**	**48.85%**	**43.22%**	**47.24%**
AVG Right			86.77%	85.21%	83.29%	76.75%	81.88%	71.83%	66.45%	82.78%	69.14%	78.88%
AVG West			**56.97%**	**53.38%**	**55.22%**	**53.16%**	**54.49%**	**48.15%**	**46.75%**	**54.64%**	**47.45%**	**52.59%**
AVG South			40.45%	43.53%	46.70%	39.00%	38.20%	39.78%	33.25%	41.58%	36.52%	40.13%
AVG East			**64.12%**	**67.59%**	**67.93%**	**58.57%**	**67.63%**	**63.82%**	**59.92%**	**65.17%**	**61.87%**	**64.23%**
AVG North			94.20%	65.07%	60.50%	57.32%	56.03%	51.41%	57.79%	66.62%	54.60%	63.19%
AVG Total lanes			**63.94%**	**57.39%**	**57.59%**	**52.01%**	**54.09%**	**50.79%**	**49.43%**	**57.00%**	**50.11%**	**55.03%**

Table 0-10 Estimation results using MLR for the second week on all lanes at junction 31616. Indicator MAPE, duration: the whole day (24 h), resolution 5 min. analysis interval: 24 h, inputs category: all the lanes at a junction and a whole week, data input: processed (smoothed) data

Lane#	Stream	Turning	MON	TUE	WED	THU	FRI	SAT	SUN	AVG Weekday	AVG Weekends	AVG all days
lane 1		Left	30.00%	47.32%	25.25%	20.06%	41.66%	22.13%	28.61%	32.86%	25.37%	30.72%
lane 2		**Left**	**39.07%**	**45.17%**	**32.76%**	**30.16%**	**44.17%**	**23.90%**	**28.38%**	**38.27%**	**26.14%**	**34.80%**
lane 3	West	Throughput	50.42%	45.33%	34.11%	39.07%	35.80%	33.55%	37.35%	40.95%	35.45%	39.38%
lane 4	stream	**Throughput**	**36.64%**	**34.34%**	**47.47%**	**33.47%**	**34.30%**	**35.48%**	**35.11%**	**37.24%**	**35.29%**	**36.69%**
lane 5		Throughput	31.15%	41.13%	34.38%	61.95%	41.34%	42.36%	39.27%	41.99%	40.81%	41.65%
lane 6		**Right**	**63.22%**	**102.65%**	**72.74%**	**70.34%**	**72.75%**	**67.90%**	**53.98%**	**76.34%**	**60.94%**	**71.94%**
lane 7		Left	23.96%	25.54%	23.09%	27.77%	21.11%	42.57%	22.78%	24.29%	32.67%	26.69%
lane 8		**Left**	**26.49%**	**25.86%**	**27.49%**	**38.05%**	**42.26%**	**30.52%**	**27.37%**	**32.03%**	**28.94%**	**31.15%**
lane 9	South	Throughput	29.00%	35.62%	27.43%	32.12%	35.32%	22.14%	19.88%	31.90%	21.01%	28.79%
lane 10	stream	**Throughput**	**16.03%**	**17.95%**	**17.85%**	**19.20%**	**19.63%**	**14.44%**	**15.10%**	**18.13%**	**14.77%**	**17.17%**
lane 11		Throughput	20.37%	20.93%	20.49%	17.59%	25.40%	15.07%	14.20%	20.96%	14.63%	19.15%
lane 12		**Right**	**85.68%**	**71.14%**	**86.06%**	**64.81%**	**72.35%**	**52.52%**	**50.07%**	**76.01%**	**51.29%**	**68.95%**
lane 13		Left	67.95%	62.86%	59.36%	53.13%	52.48%	39.73%	61.20%	59.16%	50.46%	56.67%
lane 14		**Left**	**58.50%**	**56.37%**	**49.61%**	**52.00%**	**58.31%**	**42.43%**	**57.23%**	**54.96%**	**49.83%**	**53.49%**
lane 15	East	Throughput	91.74%	65.76%	56.41%	39.55%	70.61%	59.92%	52.48%	64.81%	56.20%	62.35%
lane 16	stream	**Throughput**	**46.60%**	**41.79%**	**63.87%**	**49.73%**	**50.59%**	**49.09%**	**69.91%**	**50.52%**	**59.50%**	**53.08%**
lane 17		Throughput	37.81%	40.40%	39.05%	43.14%	41.75%	38.58%	37.59%	40.43%	38.08%	39.76%
lane 18		**Right**	**67.44%**	**88.28%**	**70.51%**	**78.50%**	**72.37%**	**72.74%**	**80.69%**	**75.42%**	**76.71%**	**75.79%**
lane 19		Left	88.90%	68.21%	59.28%	80.29%	67.75%	48.23%	48.67%	72.89%	48.45%	65.90%
lane 20		**Left**	**64.88%**	**78.49%**	**73.84%**	**54.95%**	**69.45%**	**61.30%**	**69.46%**	**68.32%**	**65.38%**	**67.48%**
lane 21	North	Throughput	251.91%	41.36%	41.27%	37.63%	17.62%	22.89%	29.38%	77.96%	26.13%	63.15%
lane 22	stream	**Throughput**	**35.20%**	**47.28%**	**40.45%**	**42.29%**	**37.58%**	**37.52%**	**51.37%**	**40.56%**	**44.44%**	**41.67%**
lane 23		Throughput	48.78%	49.62%	32.89%	23.22%	43.35%	19.85%	30.67%	39.57%	25.26%	35.48%
lane 24		**Right**	**34.61%**	**26.37%**	**32.08%**	**26.23%**	**29.76%**	**25.96%**	**25.01%**	**29.81%**	**25.49%**	**28.57%**
AVG Left			49.97%	51.23%	43.83%	44.55%	49.65%	38.85%	42.96%	47.85%	40.91%	45.86%
AVG Throughput			**57.97%**	**40.13%**	**37.97%**	**36.58%**	**37.77%**	**32.57%**	**36.03%**	**42.08%**	**34.30%**	**39.86%**
AVG Right			62.74%	72.11%	65.35%	59.97%	61.81%	54.78%	52.44%	64.39%	53.61%	61.31%
AVG West			**41.75%**	**52.66%**	**41.12%**	**42.51%**	**45.01%**	**37.55%**	**37.12%**	**44.61%**	**37.34%**	**42.53%**
AVG South			33.59%	32.84%	33.73%	33.26%	36.01%	29.54%	24.90%	33.89%	27.22%	31.98%
AVG East			**61.67%**	**59.24%**	**56.47%**	**52.67%**	**57.68%**	**50.41%**	**59.85%**	**57.55%**	**55.13%**	**56.86%**
AVG North			87.38%	51.89%	46.63%	44.10%	44.25%	35.96%	42.42%	54.85%	39.19%	50.38%
AVG Total lanes			**56.10%**	**49.16%**	**44.49%**	**43.13%**	**45.74%**	**38.37%**	**41.07%**	**47.72%**	**39.72%**	**45.44%**

Scenario 2: input range: lanes from a stream, 24h/12h/8h/4h analysis interval

Table 0-11 Estimation results using MLR for the second week on all lanes at junction 31616. Indicator MAPE, duration: the whole day (24 h), resolution 5 min. analysis interval: 24 h, inputs category: lanes from a stream and a whole week, data input: original (raw) data

Lane#	Stream	Turning	MON	TUE	WED	THU	FRI	SAT	SUN	AVG	AVG	AVG
										Weekday	Weekends	all days
lane 1		Left	25.67%	24.48%	24.11%	24.20%	30.39%	22.73%	28.56%	25.77%	25.65%	25.73%
lane 2		**Left**	**25.63%**	**34.17%**	**28.71%**	**28.63%**	**24.67%**	**23.70%**	**27.63%**	**28.36%**	**25.67%**	**27.59%**
lane 3	West	Throughput	40.69%	37.29%	36.62%	32.06%	30.87%	27.83%	32.00%	35.51%	29.92%	33.91%
lane 4	stream	**Throughput**	**31.74%**	**31.39%**	**29.48%**	**28.94%**	**24.06%**	**25.33%**	**23.87%**	**29.12%**	**24.60%**	**27.83%**
lane 5		Throughput	27.59%	28.19%	29.17%	31.13%	28.49%	27.32%	31.02%	28.91%	29.17%	28.99%
lane 6		**Right**	**86.94%**	**67.81%**	**69.96%**	**61.54%**	**66.12%**	**58.30%**	**58.08%**	**70.47%**	**58.19%**	**66.96%**
lane 7		Left	19.13%	29.28%	32.72%	22.97%	21.62%	24.93%	21.91%	25.14%	23.42%	24.65%
lane 8		**Left**	**24.48%**	**23.41%**	**29.83%**	**24.62%**	**26.57%**	**24.39%**	**19.52%**	**25.78%**	**21.96%**	**24.69%**
lane 9	South	Throughput	22.93%	29.93%	25.76%	25.27%	28.96%	25.44%	25.89%	26.57%	25.67%	26.31%
lane 10	stream	**Throughput**	**15.10%**	**19.83%**	**17.60%**	**21.25%**	**15.34%**	**15.92%**	**14.13%**	**17.82%**	**15.03%**	**17.02%**
lane 11		Throughput	20.72%	16.58%	19.69%	19.98%	35.10%	15.71%	14.51%	22.41%	15.11%	20.33%
lane 12		**Right**	**56.78%**	**52.38%**	**58.38%**	**49.46%**	**62.45%**	**45.46%**	**40.02%**	**55.89%**	**42.74%**	**52.13%**
lane 13		Left	44.82%	40.09%	47.11%	44.58%	35.35%	40.37%	37.48%	42.39%	38.93%	41.40%
lane 14		**Left**	**40.52%**	**42.52%**	**39.79%**	**42.22%**	**34.72%**	**39.56%**	**34.73%**	**39.95%**	**37.15%**	**39.15%**
lane 15	East	Throughput	54.23%	42.77%	39.60%	33.04%	44.09%	39.90%	36.00%	42.75%	37.95%	41.38%
lane 16	stream	**Throughput**	**41.20%**	**41.42%**	**45.29%**	**44.87%**	**43.18%**	**40.79%**	**32.80%**	**43.19%**	**36.80%**	**41.36%**
lane 17		Throughput	28.48%	29.28%	31.02%	25.55%	32.69%	30.10%	31.26%	29.40%	30.68%	29.77%
lane 18		**Right**	**66.71%**	**64.79%**	**56.94%**	**60.01%**	**64.75%**	**58.29%**	**50.88%**	**62.64%**	**54.59%**	**60.34%**
lane 19		Left	49.15%	53.07%	49.81%	41.89%	58.98%	46.73%	49.60%	50.58%	48.17%	49.89%
lane 20		**Left**	**52.43%**	**66.43%**	**79.78%**	**61.31%**	**49.65%**	**51.76%**	**42.72%**	**61.92%**	**47.24%**	**57.73%**
lane 21	North	Throughput	259.76%	31.64%	27.18%	25.54%	23.55%	21.03%	51.84%	73.53%	36.44%	62.93%
lane 22	stream	**Throughput**	**41.40%**	**36.94%**	**44.80%**	**47.07%**	**42.13%**	**49.01%**	**44.84%**	**42.47%**	**46.93%**	**43.74%**
lane 23		Throughput	35.66%	27.63%	20.61%	25.24%	43.15%	33.00%	49.24%	30.46%	41.12%	33.50%
lane 24		**Right**	**24.69%**	**35.11%**	**31.34%**	**27.18%**	**34.69%**	**28.58%**	**27.22%**	**30.60%**	**27.90%**	**29.83%**
AVG Left			35.23%	39.18%	41.48%	36.30%	35.24%	34.27%	32.77%	37.49%	33.52%	36.35%
AVG Throughput			**51.63%**	**31.07%**	**30.57%**	**30.00%**	**32.63%**	**29.28%**	**32.28%**	**35.18%**	**30.78%**	**33.92%**
AVG Right			58.78%	55.02%	54.16%	49.55%	57.00%	47.66%	44.05%	54.90%	45.85%	52.32%
AVG West			**39.71%**	**37.22%**	**36.34%**	**34.42%**	**34.10%**	**30.87%**	**33.53%**	**36.36%**	**32.20%**	**35.17%**
AVG South			26.52%	28.57%	30.66%	27.26%	31.67%	25.31%	22.66%	28.94%	23.99%	27.52%
AVG East			**45.99%**	**43.48%**	**43.29%**	**41.71%**	**42.46%**	**41.50%**	**37.19%**	**43.39%**	**39.35%**	**42.23%**
AVG North			77.18%	41.80%	42.25%	38.04%	42.03%	38.35%	44.24%	48.26%	41.30%	46.27%
AVG Total lanes			**47.35%**	**37.77%**	**38.14%**	**35.36%**	**37.57%**	**34.01%**	**34.41%**	**39.24%**	**34.21%**	**37.80%**

Table 0-12 Estimation results using MLR for the second week on all lanes at junction 31616. Indicator MAPE, duration: the whole day (24 h), resolution 5 min. analysis interval: 24 h, inputs category: lanes from a stream and a whole week, processed (smoothed) data

Lane#	Stream	Turning	MON	TUE	WED	THU	FRI	SAT	SUN	AVG	AVG	AVG
										Weekday	Weekends	all days
lane 1		Left	19.02%	16.66%	19.98%	13.45%	22.06%	14.43%	20.77%	18.23%	17.60%	18.05%
lane 2		**Left**	**18.76%**	**23.19%**	**21.44%**	**19.86%**	**19.02%**	**16.11%**	**18.70%**	**20.45%**	**17.41%**	**19.58%**
lane 3	West	Throughput	27.06%	30.21%	27.80%	20.82%	25.64%	20.64%	22.29%	26.31%	21.47%	24.92%
lane 4	stream	**Throughput**	**25.09%**	**23.93%**	**19.98%**	**18.58%**	**18.32%**	**14.12%**	**17.58%**	**21.18%**	**15.85%**	**19.66%**
lane 5		Throughput	16.98%	20.51%	18.11%	21.43%	24.11%	22.16%	26.02%	20.23%	24.09%	21.33%
lane 6		**Right**	**55.12%**	**63.91%**	**64.40%**	**50.33%**	**41.80%**	**41.44%**	**35.35%**	**55.11%**	**38.40%**	**50.34%**
lane 7		Left	12.95%	20.19%	15.66%	16.05%	9.87%	26.23%	13.30%	14.94%	19.77%	16.32%
lane 8		**Left**	**13.21%**	**18.54%**	**15.21%**	**21.26%**	**18.11%**	**17.59%**	**11.93%**	**17.27%**	**14.76%**	**16.55%**
lane 9	South	Throughput	15.46%	19.64%	16.36%	17.09%	25.54%	15.71%	14.94%	18.82%	15.33%	17.82%
lane 10	stream	**Throughput**	**11.72%**	**11.84%**	**11.61%**	**13.54%**	**10.18%**	**9.49%**	**8.63%**	**11.78%**	**9.06%**	**11.00%**
lane 11		Throughput	13.73%	11.54%	12.60%	11.47%	28.69%	9.03%	11.35%	15.61%	10.19%	14.06%
lane 12		**Right**	**35.52%**	**50.22%**	**43.07%**	**35.38%**	**43.23%**	**35.85%**	**29.66%**	**41.48%**	**32.76%**	**38.99%**
lane 13		Left	27.12%	28.71%	27.25%	29.88%	29.36%	28.53%	24.67%	28.46%	26.60%	27.93%
lane 14		**Left**	**31.55%**	**30.48%**	**28.01%**	**30.40%**	**28.82%**	**27.58%**	**24.24%**	**29.85%**	**25.91%**	**28.73%**
lane 15	East	Throughput	47.79%	38.36%	33.79%	29.05%	46.39%	36.47%	33.99%	39.08%	35.23%	37.98%
lane 16	stream	**Throughput**	**29.55%**	**24.16%**	**29.19%**	**25.51%**	**31.20%**	**25.45%**	**22.99%**	**27.92%**	**24.22%**	**26.86%**
lane 17		Throughput	18.01%	20.25%	23.51%	22.42%	25.74%	21.81%	21.36%	21.99%	21.59%	21.87%
lane 18		**Right**	**44.55%**	**40.95%**	**41.97%**	**44.83%**	**48.95%**	**45.69%**	**38.46%**	**44.25%**	**42.08%**	**43.63%**
lane 19		Left	55.70%	45.12%	37.66%	45.63%	47.64%	41.12%	39.53%	46.35%	40.33%	44.63%
lane 20		**Left**	**39.82%**	**41.69%**	**46.59%**	**39.58%**	**38.56%**	**38.90%**	**35.11%**	**41.25%**	**37.01%**	**40.04%**
lane 21	North	Throughput	267.59%	32.66%	19.53%	24.17%	16.67%	18.54%	26.21%	72.12%	22.38%	57.91%
lane 22	stream	**Throughput**	**29.66%**	**29.05%**	**23.16%**	**31.09%**	**27.10%**	**28.59%**	**30.46%**	**28.01%**	**29.53%**	**28.44%**
lane 23		Throughput	28.81%	23.92%	16.24%	17.66%	37.50%	18.23%	26.48%	24.83%	22.36%	24.12%
lane 24		**Right**	**19.09%**	**23.52%**	**18.77%**	**15.42%**	**21.90%**	**18.44%**	**20.37%**	**19.74%**	**19.41%**	**19.64%**
AVG Left			27.27%	28.07%	26.48%	27.01%	26.68%	26.31%	23.53%	27.10%	24.92%	26.48%
AVG Throughput			**44.29%**	**23.84%**	**20.99%**	**21.07%**	**26.42%**	**20.02%**	**21.86%**	**27.32%**	**20.94%**	**25.50%**
AVG Right			38.57%	44.65%	42.05%	36.49%	38.97%	35.36%	30.96%	40.15%	33.16%	38.15%
AVG West			**27.01%**	**29.74%**	**28.62%**	**24.08%**	**25.16%**	**21.48%**	**23.45%**	**26.92%**	**22.47%**	**25.65%**
AVG South			17.10%	22.00%	19.09%	19.13%	22.60%	18.98%	14.97%	19.98%	16.98%	19.12%
AVG East			**33.10%**	**30.49%**	**30.62%**	**30.35%**	**35.08%**	**30.92%**	**27.62%**	**31.93%**	**29.27%**	**31.17%**
AVG North			73.45%	32.66%	26.99%	28.93%	31.56%	27.30%	29.69%	38.72%	28.50%	35.80%
AVG Total lanes			**37.66%**	**28.72%**	**26.33%**	**25.62%**	**28.60%**	**24.67%**	**23.93%**	**29.39%**	**24.30%**	**27.93%**

Table 0-13 Estimation results using MLR for the second week on all lanes at junction 31616. Indicator MAPE, duration: the whole day (24 h), resolution 5 min. analysis interval: 12h, inputs category: lanes from a stream and a whole week, data input: original (raw) data

Lane#	Stream	Turning	MON	TUE	WED	THU	FRI	SAT	SUN	AVG Weekday	AVG Weekends	AVG all days
lane 1		Left	30.70%	26.93%	24.65%	24.39%	31.87%	22.95%	29.42%	27.71%	26.19%	27.27%
lane 2		**Left**	**32.00%**	**39.41%**	**30.08%**	**34.14%**	**30.99%**	**25.18%**	**28.86%**	**33.32%**	**27.02%**	**31.52%**
lane 3	West	Throughput	47.25%	40.33%	37.30%	37.25%	37.25%	30.08%	33.78%	39.88%	31.93%	37.61%
lane 4	stream	**Throughput**	**37.13%**	**34.95%**	**31.99%**	**31.61%**	**25.94%**	**26.07%**	**26.69%**	**32.32%**	**26.38%**	**30.63%**
lane 5		Throughput	30.18%	32.78%	32.43%	34.44%	32.15%	29.29%	33.66%	32.40%	31.48%	32.13%
lane 6		**Right**	**88.81%**	**73.03%**	**80.78%**	**77.50%**	**76.31%**	**66.97%**	**67.06%**	**79.29%**	**67.02%**	**75.78%**
lane 7		Left	19.95%	28.77%	35.00%	26.11%	23.05%	28.43%	23.83%	26.58%	26.13%	26.45%
lane 8		**Left**	**28.54%**	**28.41%**	**28.48%**	**27.49%**	**30.06%**	**29.71%**	**20.53%**	**28.60%**	**25.12%**	**27.60%**
lane 9	South	Throughput	28.86%	36.82%	30.46%	26.40%	29.16%	26.63%	29.67%	30.34%	28.15%	29.71%
lane 10	stream	**Throughput**	**16.33%**	**21.73%**	**21.62%**	**20.23%**	**17.39%**	**16.61%**	**16.47%**	**19.46%**	**16.54%**	**18.63%**
lane 11		Throughput	21.22%	18.09%	21.73%	21.30%	32.68%	15.87%	16.73%	23.00%	16.30%	21.09%
lane 12		**Right**	**74.08%**	**66.80%**	**64.73%**	**57.77%**	**61.52%**	**51.16%**	**55.40%**	**64.98%**	**53.28%**	**61.64%**
lane 13		Left	46.76%	44.21%	50.46%	46.13%	40.27%	44.41%	40.17%	45.57%	42.29%	44.63%
lane 14		**Left**	**45.17%**	**46.07%**	**40.56%**	**44.48%**	**37.71%**	**43.89%**	**39.00%**	**42.80%**	**41.45%**	**42.41%**
lane 15	East	Throughput	59.30%	50.75%	41.73%	37.66%	46.07%	45.68%	47.64%	47.10%	46.66%	46.98%
lane 16	stream	**Throughput**	**51.79%**	**43.52%**	**49.50%**	**50.92%**	**49.38%**	**46.41%**	**38.99%**	**49.02%**	**42.70%**	**47.22%**
lane 17		Throughput	33.56%	34.30%	34.07%	28.78%	37.14%	33.17%	33.51%	33.57%	33.34%	33.50%
lane 18		**Right**	**74.66%**	**67.07%**	**65.54%**	**65.72%**	**66.68%**	**64.10%**	**49.77%**	**67.93%**	**56.94%**	**64.79%**
lane 19		Left	49.77%	61.30%	50.35%	48.63%	64.99%	46.73%	54.06%	55.01%	50.40%	53.69%
lane 20		**Left**	**58.47%**	**78.62%**	**82.45%**	**75.19%**	**52.68%**	**52.39%**	**48.73%**	**69.48%**	**50.56%**	**64.08%**
lane 21	North	Throughput	268.11%	36.25%	28.89%	33.83%	25.20%	22.39%	49.75%	78.46%	36.07%	66.35%
lane 22	stream	**Throughput**	**42.32%**	**43.30%**	**51.77%**	**52.31%**	**43.62%**	**43.00%**	**47.61%**	**46.66%**	**45.31%**	**46.28%**
lane 23		Throughput	40.92%	29.84%	27.29%	28.43%	42.25%	34.89%	52.35%	33.75%	43.62%	36.57%
lane 24		**Right**	**27.89%**	**35.98%**	**32.98%**	**31.09%**	**35.41%**	**30.21%**	**27.94%**	**32.67%**	**29.08%**	**31.64%**
AVG Left			38.92%	44.22%	42.75%	40.82%	38.95%	36.71%	35.58%	41.13%	36.14%	39.71%
AVG Throughput			**56.41%**	**35.22%**	**34.07%**	**33.60%**	**34.85%**	**30.84%**	**35.57%**	**38.83%**	**33.21%**	**37.22%**
AVG Right			66.36%	60.72%	61.01%	58.02%	59.98%	53.11%	50.04%	61.22%	51.58%	58.46%
AVG West			**44.35%**	**41.24%**	**39.54%**	**39.89%**	**39.09%**	**33.42%**	**36.58%**	**40.82%**	**35.00%**	**39.16%**
AVG South			31.50%	33.44%	33.67%	29.88%	32.31%	28.07%	27.11%	32.16%	27.59%	30.85%
AVG East			**51.87%**	**47.65%**	**46.98%**	**45.62%**	**46.21%**	**46.28%**	**41.51%**	**47.67%**	**43.90%**	**46.59%**
AVG North			81.25%	47.55%	45.62%	44.91%	44.03%	38.27%	46.74%	52.67%	42.50%	49.77%
AVG Total lanes			**52.24%**	**42.47%**	**41.45%**	**40.08%**	**40.41%**	**36.51%**	**37.98%**	**43.33%**	**37.25%**	**41.59%**

Table 0-14 Estimation results using MLR for the second week on all lanes at junction 31616. Indicator MAPE, duration: the whole day (24 h), resolution 5 min. analysis interval: 12 h, inputs category: lanes from a stream and a whole week, processed (smoothed) data

Lane#	Stream	Turning	MON	TUE	WED	THU	FRI	SAT	SUN	AVG	AVG	AVG
										Weekday	Weekends	all days
lane 1		Left	14.93%	15.73%	14.85%	13.69%	17.96%	14.39%	19.44%	15.43%	16.92%	15.86%
lane 2		**Left**	**20.04%**	**24.23%**	**20.36%**	**20.75%**	**20.72%**	**17.20%**	**18.18%**	**21.22%**	**17.69%**	**20.21%**
lane 3	West	Throughput	29.69%	27.34%	25.69%	23.53%	25.80%	20.62%	23.61%	26.41%	22.12%	25.18%
lane 4	stream	**Throughput**	**23.09%**	**22.43%**	**20.91%**	**17.87%**	**19.14%**	**15.20%**	**14.84%**	**20.69%**	**15.02%**	**19.07%**
lane 5		Throughput	18.77%	19.50%	18.92%	20.90%	20.68%	21.93%	21.76%	19.75%	21.85%	20.35%
lane 6		**Right**	**52.14%**	**61.04%**	**55.66%**	**45.61%**	**41.67%**	**41.14%**	**40.56%**	**51.22%**	**40.85%**	**48.26%**
lane 7		Left	12.22%	20.33%	15.20%	15.06%	12.51%	16.07%	14.86%	15.06%	15.47%	15.18%
lane 8		**Left**	**16.18%**	**15.48%**	**17.51%**	**14.87%**	**20.53%**	**14.89%**	**13.64%**	**16.91%**	**14.27%**	**16.16%**
lane 9	South	Throughput	18.83%	23.26%	20.09%	17.42%	20.68%	14.07%	14.49%	20.06%	14.28%	18.41%
lane 10	stream	**Throughput**	**11.49%**	**11.28%**	**13.13%**	**11.58%**	**10.44%**	**8.93%**	**9.55%**	**11.58%**	**9.24%**	**10.91%**
lane 11		Throughput	12.84%	11.83%	14.46%	9.93%	24.44%	9.10%	11.22%	14.70%	10.16%	13.40%
lane 12		**Right**	**47.96%**	**48.43%**	**39.14%**	**36.11%**	**42.63%**	**34.53%**	**35.50%**	**42.85%**	**35.02%**	**40.61%**
lane 13		Left	30.85%	31.21%	31.48%	29.32%	30.79%	29.58%	28.96%	30.73%	29.27%	30.31%
lane 14		**Left**	**29.54%**	**33.10%**	**28.96%**	**30.07%**	**28.41%**	**29.93%**	**26.19%**	**30.02%**	**28.06%**	**29.46%**
lane 15	East	Throughput	47.58%	42.94%	33.67%	32.21%	45.22%	36.93%	35.06%	40.32%	36.00%	39.09%
lane 16	stream	**Throughput**	**30.79%**	**26.56%**	**33.25%**	**31.04%**	**31.46%**	**27.96%**	**23.12%**	**30.62%**	**25.54%**	**29.17%**
lane 17		Throughput	17.55%	24.12%	23.06%	24.07%	23.44%	22.81%	21.25%	22.45%	22.03%	22.33%
lane 18		**Right**	**52.11%**	**45.77%**	**44.01%**	**50.55%**	**44.62%**	**41.97%**	**38.44%**	**47.41%**	**40.21%**	**45.35%**
lane 19		Left	39.88%	47.30%	32.00%	35.67%	50.79%	39.33%	37.57%	41.13%	38.45%	40.36%
lane 20		**Left**	**44.17%**	**62.85%**	**48.56%**	**41.97%**	**35.48%**	**35.65%**	**30.22%**	**46.61%**	**32.94%**	**42.70%**
lane 21	North	Throughput	265.66%	22.40%	19.63%	24.46%	15.38%	17.36%	25.64%	69.51%	21.50%	55.79%
lane 22	stream	**Throughput**	**31.54%**	**31.30%**	**29.63%**	**27.96%**	**25.41%**	**25.34%**	**30.00%**	**29.17%**	**27.67%**	**28.74%**
lane 23		Throughput	29.54%	17.15%	18.35%	16.63%	37.09%	17.50%	26.73%	23.75%	22.12%	23.28%
lane 24		**Right**	**16.48%**	**22.47%**	**20.22%**	**20.03%**	**19.86%**	**17.01%**	**19.04%**	**19.81%**	**18.03%**	**19.30%**
AVG Left			25.98%	31.28%	26.12%	25.18%	27.15%	24.63%	23.63%	27.14%	24.13%	26.28%
AVG Throughput			**44.78%**	**23.34%**	**22.57%**	**21.47%**	**24.93%**	**19.81%**	**21.44%**	**27.42%**	**20.63%**	**25.48%**
AVG Right			42.17%	44.43%	39.76%	38.08%	37.20%	33.66%	33.39%	40.33%	33.52%	38.38%
AVG West			**26.44%**	**28.38%**	**26.07%**	**23.73%**	**24.33%**	**21.75%**	**23.07%**	**25.79%**	**22.41%**	**24.82%**
AVG South			19.92%	21.77%	19.92%	17.50%	21.87%	16.27%	16.54%	20.20%	16.40%	19.11%
AVG East			**34.74%**	**33.95%**	**32.41%**	**32.88%**	**33.99%**	**31.53%**	**28.84%**	**33.59%**	**30.18%**	**32.62%**
AVG North			71.21%	33.91%	28.07%	27.79%	30.67%	25.37%	28.20%	38.33%	26.78%	35.03%
AVG Total lanes			**38.08%**	**29.50%**	**26.61%**	**25.47%**	**27.71%**	**23.73%**	**24.16%**	**29.48%**	**23.94%**	**27.90%**

Table 0-15 Estimation results using MLR for the second week on all lanes at junction 31616. Indicator MAPE, duration: the whole day (24 h), resolution 5 min. analysis interval: 8h, inputs category: lanes from a stream and a whole week, data input: original (raw) data

Lane#	Stream	Turning	MON	TUE	WED	THU	FRI	SAT	SUN	AVG Weekday	AVG Weekends	AVG all days
lane 1		Left	31.21%	32.92%	31.63%	29.07%	36.80%	30.25%	34.92%	32.33%	32.59%	32.40%
lane 2		Left	40.51%	50.42%	34.47%	39.08%	36.21%	35.54%	35.39%	40.14%	35.47%	38.80%
lane 3	West	Throughput	54.55%	47.85%	55.10%	48.93%	45.58%	38.22%	36.44%	50.40%	37.33%	46.67%
lane 4	stream	Throughput	43.12%	47.48%	41.87%	39.82%	30.47%	37.42%	35.75%	40.55%	36.59%	39.42%
lane 5		Throughput	32.97%	40.59%	44.79%	40.64%	43.35%	38.03%	36.75%	40.47%	37.39%	39.59%
lane 6		Right	97.87%	81.73%	93.62%	90.69%	86.30%	71.43%	72.46%	90.04%	71.95%	84.87%
lane 7		Left	24.75%	37.09%	33.94%	30.16%	28.40%	28.31%	27.39%	30.87%	27.85%	30.01%
lane 8		Left	34.36%	34.33%	31.18%	33.43%	39.34%	31.89%	28.07%	34.53%	29.98%	33.23%
lane 9	South	Throughput	29.68%	45.92%	35.39%	35.50%	29.63%	30.97%	26.73%	35.22%	28.85%	33.40%
lane 10	stream	Throughput	18.55%	25.59%	23.00%	22.93%	23.53%	20.59%	18.42%	22.72%	19.51%	21.80%
lane 11		Throughput	24.47%	21.37%	23.80%	24.34%	29.75%	21.95%	17.07%	24.75%	19.51%	23.25%
lane 12		Right	71.16%	61.24%	70.19%	66.32%	65.25%	60.39%	48.89%	66.83%	54.64%	63.35%
lane 13		Left	53.41%	47.20%	54.97%	45.82%	45.28%	47.36%	47.96%	49.34%	47.66%	48.86%
lane 14		Left	48.50%	45.78%	46.17%	46.76%	46.21%	48.46%	44.03%	46.68%	46.25%	46.56%
lane 15	East	Throughput	66.94%	50.49%	46.67%	41.49%	47.60%	49.88%	48.00%	50.64%	48.94%	50.15%
lane 16	stream	Throughput	52.05%	43.15%	48.85%	58.96%	50.55%	49.40%	44.27%	50.71%	46.84%	49.60%
lane 17		Throughput	36.86%	43.12%	38.13%	35.08%	42.42%	34.97%	37.49%	39.12%	36.23%	38.30%
lane 18		Right	77.57%	80.06%	66.41%	85.19%	71.43%	69.70%	59.54%	76.13%	64.62%	72.84%
lane 19		Left	64.52%	65.03%	57.35%	61.83%	64.92%	55.29%	62.64%	62.73%	58.97%	61.65%
lane 20		Left	59.39%	80.46%	88.78%	88.76%	59.81%	61.05%	58.23%	75.44%	59.64%	70.93%
lane 21	North	Throughput	277.47%	44.56%	30.51%	33.96%	28.27%	24.83%	55.73%	82.95%	40.28%	70.76%
lane 22	stream	Throughput	57.05%	41.41%	53.84%	54.70%	48.46%	56.59%	47.76%	51.09%	52.18%	51.40%
lane 23		Throughput	57.11%	30.18%	32.43%	27.10%	46.81%	38.99%	54.06%	38.73%	46.53%	40.95%
lane 24		Right	33.65%	36.19%	37.97%	33.72%	37.92%	36.79%	32.50%	35.89%	34.65%	35.53%
AVG Left			44.58%	49.15%	47.31%	46.86%	44.62%	42.27%	42.33%	46.51%	42.30%	45.30%
AVG Throughput			62.57%	40.14%	39.53%	38.62%	38.87%	36.82%	38.21%	43.95%	37.51%	42.11%
AVG Right			70.06%	64.81%	67.05%	68.98%	65.23%	59.58%	53.35%	67.22%	56.46%	64.15%
AVG West			50.04%	50.17%	50.25%	48.04%	46.45%	41.82%	41.95%	48.99%	41.88%	46.96%
AVG South			33.83%	37.59%	36.25%	35.45%	35.98%	32.35%	27.76%	35.82%	30.06%	34.17%
AVG East			55.89%	51.63%	50.20%	52.22%	50.58%	49.96%	46.88%	52.10%	48.42%	51.05%
AVG North			91.53%	49.64%	50.15%	50.01%	47.70%	45.59%	51.82%	57.81%	48.71%	55.21%
AVG Total lanes			57.82%	47.26%	46.71%	46.43%	45.18%	42.43%	42.10%	48.68%	42.27%	46.85%

Table 0-16 Estimation results using MLR for the second week on all lanes at junction 31616. Indicator MAPE, duration: the whole day (24 h), resolution 5 min. analysis interval: 8h, inputs category: lanes from a stream and a whole week, data input: processed (smoothed) data

Lane#	Stream	Turning	MON	TUE	WED	THU	FRI	SAT	SUN	AVG Weekday	AVG Weekends	AVG all days
lane 1		Left	16.20%	19.60%	20.11%	17.80%	21.77%	17.73%	21.74%	19.10%	19.74%	19.28%
lane 2		Left	23.09%	31.40%	23.75%	21.56%	22.59%	21.45%	21.43%	24.48%	21.44%	23.61%
lane 3	West	Throughput	41.74%	28.39%	38.98%	30.62%	27.23%	24.23%	24.76%	33.39%	24.50%	30.85%
lane 4	stream	Throughput	27.07%	31.49%	29.10%	23.79%	20.83%	24.49%	21.35%	26.46%	22.92%	25.45%
lane 5		Throughput	20.62%	29.10%	25.41%	23.42%	26.12%	23.30%	25.51%	24.93%	24.41%	24.78%
lane 6		Right	60.04%	58.09%	60.11%	51.93%	47.64%	40.93%	46.95%	55.56%	43.94%	52.24%
lane 7		Left	16.28%	25.60%	16.67%	18.14%	15.51%	14.32%	17.75%	18.44%	16.04%	17.75%
lane 8		Left	20.73%	20.38%	18.19%	18.69%	26.98%	18.17%	15.03%	20.99%	16.60%	19.74%
lane 9	South	Throughput	21.22%	34.62%	25.94%	26.07%	22.10%	16.58%	14.48%	25.99%	15.53%	23.00%
lane 10	stream	Throughput	11.46%	14.22%	14.31%	13.58%	14.25%	11.43%	11.12%	13.56%	11.28%	12.91%
lane 11		Throughput	16.14%	15.85%	15.22%	13.47%	20.52%	12.98%	10.72%	16.24%	11.85%	14.99%
lane 12		Right	50.22%	43.53%	40.82%	39.85%	43.76%	38.86%	38.10%	43.64%	38.48%	42.16%
lane 13		Left	38.04%	33.32%	35.21%	29.38%	33.16%	32.04%	32.84%	33.82%	32.44%	33.43%
lane 14		Left	36.21%	34.79%	31.02%	30.01%	33.75%	36.51%	29.12%	33.16%	32.82%	33.06%
lane 15	East	Throughput	52.89%	41.04%	35.66%	31.43%	40.69%	35.03%	36.45%	40.34%	35.74%	39.03%
lane 16	stream	Throughput	33.25%	27.28%	32.18%	37.93%	36.48%	29.99%	27.29%	33.42%	28.64%	32.06%
lane 17		Throughput	21.33%	28.40%	24.28%	25.02%	26.49%	23.42%	26.13%	25.10%	24.78%	25.01%
lane 18		Right	50.15%	51.66%	43.45%	58.77%	46.67%	49.59%	44.03%	50.14%	46.81%	49.19%
lane 19		Left	54.48%	50.43%	35.83%	47.28%	47.27%	45.64%	41.58%	47.06%	43.61%	46.07%
lane 20		Left	44.86%	58.09%	49.19%	50.86%	41.59%	42.22%	38.95%	48.92%	40.59%	46.54%
lane 21	North	Throughput	278.65%	29.45%	19.60%	23.78%	17.98%	19.18%	30.11%	73.89%	24.65%	59.82%
lane 22	stream	Throughput	35.28%	27.97%	31.14%	30.74%	26.14%	33.85%	32.62%	30.25%	33.24%	31.11%
lane 23		Throughput	33.99%	17.04%	23.00%	15.12%	39.03%	18.95%	29.99%	25.64%	24.47%	25.30%
lane 24		Right	21.89%	21.79%	22.75%	22.37%	23.67%	24.31%	21.91%	22.49%	23.11%	22.67%
AVG Left			31.24%	34.20%	28.75%	29.22%	30.33%	28.51%	27.31%	30.75%	27.91%	29.93%
AVG Throughput			49.47%	27.07%	26.24%	24.58%	26.49%	22.79%	24.21%	30.77%	23.50%	28.69%
AVG Right			45.58%	43.77%	41.78%	43.23%	40.44%	38.42%	37.75%	42.96%	38.09%	41.57%
AVG West			31.46%	33.01%	32.91%	28.19%	27.70%	25.36%	26.96%	30.65%	26.16%	29.37%
AVG South			22.68%	25.70%	21.86%	21.63%	23.85%	18.72%	17.87%	23.14%	18.30%	21.76%
AVG East			38.65%	36.08%	33.63%	35.42%	36.21%	34.43%	32.64%	36.00%	33.54%	35.29%
AVG North			78.19%	34.13%	30.25%	31.69%	32.61%	30.69%	32.53%	41.38%	31.61%	38.59%
AVG Total lanes			42.74%	32.23%	29.66%	29.23%	30.09%	27.30%	27.50%	32.79%	27.40%	31.25%

Table 0-17 Estimation results using MLR for the second week on all lanes at junction 31616. Indicator MAPE, duration: the whole day (24 h), resolution 5 min. analysis interval: 4h, inputs category: lanes from a stream and a whole week, data input: original (raw) data

Lane#	Stream	Turning	MON	TUE	WED	THU	FRI	SAT	SUN	AVG Weekday	AVG Weekends	AVG all days
lane 1		Left	71.55%	69.34%	89.77%	63.56%	83.38%	66.07%	72.19%	75.52%	69.13%	73.69%
lane 2		Left	74.88%	98.42%	72.93%	97.69%	86.54%	88.23%	88.13%	86.09%	88.18%	86.69%
lane 3	West	Throughput	103.48%	101.56%	90.21%	99.42%	97.06%	89.03%	98.17%	98.35%	93.60%	96.99%
lane 4	stream	Throughput	73.94%	82.41%	82.31%	89.76%	85.29%	93.74%	67.59%	82.74%	80.67%	82.15%
lane 5		Throughput	104.38%	90.38%	105.15%	74.26%	86.68%	84.41%	82.25%	92.17%	83.33%	89.64%
lane 6		Right	200.38%	227.03%	202.22%	187.82%	233.26%	149.88%	167.89%	210.14%	158.89%	195.50%
lane 7		Left	63.71%	65.92%	91.85%	60.07%	58.94%	65.26%	56.86%	68.10%	61.06%	66.09%
lane 8		Left	71.75%	82.59%	53.48%	78.80%	83.60%	66.93%	54.37%	74.04%	60.65%	70.22%
lane 9	South	Throughput	53.24%	97.53%	70.53%	91.17%	52.68%	69.06%	49.59%	73.03%	59.33%	69.11%
lane 10	stream	Throughput	40.82%	54.45%	45.76%	62.72%	44.07%	37.64%	44.02%	49.56%	40.83%	47.07%
lane 11		Throughput	48.55%	54.42%	48.97%	44.78%	53.53%	48.62%	37.01%	50.05%	42.82%	47.98%
lane 12		Right	181.25%	142.10%	173.58%	121.99%	134.49%	135.24%	100.87%	150.68%	118.06%	141.36%
lane 13		Left	124.37%	95.47%	121.53%	102.67%	108.91%	91.66%	89.50%	110.59%	90.58%	104.87%
lane 14		Left	105.97%	95.25%	102.35%	89.80%	108.50%	81.46%	92.44%	100.37%	86.95%	96.54%
lane 15	East	Throughput	134.09%	111.73%	122.68%	87.14%	99.75%	94.60%	74.87%	111.08%	84.74%	103.55%
lane 16	stream	Throughput	122.34%	104.35%	119.82%	111.73%	127.32%	98.79%	93.25%	117.11%	96.02%	111.09%
lane 17		Throughput	73.56%	75.01%	78.63%	89.30%	89.40%	84.65%	89.40%	81.18%	87.03%	82.85%
lane 18		Right	168.46%	144.35%	152.50%	160.34%	146.13%	138.22%	142.15%	154.36%	140.19%	150.31%
lane 19		Left	172.91%	193.43%	154.93%	171.39%	145.17%	106.20%	131.10%	167.57%	118.65%	153.59%
lane 20		Left	133.44%	197.15%	193.19%	191.88%	132.42%	176.26%	120.29%	169.62%	148.28%	163.52%
lane 21	North	Throughput	337.85%	104.56%	94.41%	99.86%	70.39%	64.44%	92.59%	141.41%	78.52%	123.44%
lane 22	stream	Throughput	103.31%	92.92%	146.65%	118.82%	103.46%	110.31%	102.11%	113.03%	106.21%	111.08%
lane 23		Throughput	110.16%	102.63%	99.41%	57.17%	94.84%	83.09%	79.17%	92.84%	81.13%	89.50%
lane 24		Right	98.10%	84.01%	81.70%	98.23%	68.45%	81.06%	76.87%	86.10%	78.97%	84.06%
AVG Left			102.32%	112.20%	110.00%	106.98%	100.93%	92.76%	88.11%	106.49%	90.43%	101.90%
AVG Throughput			108.81%	89.33%	92.04%	85.51%	83.71%	79.87%	75.84%	91.88%	77.85%	87.87%
AVG Right			162.05%	149.37%	152.50%	142.10%	145.58%	126.10%	121.95%	150.32%	124.02%	142.81%
AVG West			104.77%	111.52%	107.10%	102.09%	112.04%	95.23%	96.04%	107.50%	95.63%	104.11%
AVG South			76.55%	82.84%	80.70%	76.59%	71.22%	70.46%	57.12%	77.58%	63.79%	73.64%
AVG East			121.47%	104.36%	116.25%	106.83%	113.34%	98.23%	96.94%	112.45%	97.58%	108.20%
AVG North			159.30%	129.12%	128.38%	122.89%	102.46%	103.56%	100.36%	128.43%	101.96%	120.87%
AVG Total lanes			115.52%	106.96%	108.11%	102.10%	99.76%	91.87%	87.61%	106.49%	89.74%	101.70%

Table 0-18 Estimation results using MLR for the second week on all lanes at junction 31616. Indicator MAPE, duration: the whole day (24 h), resolution 5 min. analysis interval: 4h, inputs category: lanes from a stream and a whole week, data input: processed (smoothed) data

Lane#	Stream	Turning	MON	TUE	WED	THU	FRI	SAT	SUN	AVG Weekday	AVG Weekends	AVG all days
lane 1		Left	43.65%	45.14%	59.11%	39.32%	58.03%	42.13%	46.82%	49.05%	44.48%	47.74%
lane 2		Left	49.41%	63.52%	53.30%	59.97%	57.79%	62.69%	58.07%	56.80%	60.38%	57.82%
lane 3	West	Throughput	78.64%	66.53%	64.70%	72.13%	65.07%	64.59%	67.12%	69.41%	65.86%	68.40%
lane 4	stream	Throughput	45.93%	56.73%	52.13%	59.86%	59.19%	65.84%	44.68%	54.77%	55.26%	54.91%
lane 5		Throughput	70.50%	65.50%	74.53%	48.88%	57.52%	56.78%	57.85%	63.39%	57.32%	61.65%
lane 6		Right	111.53%	142.70%	141.89%	110.25%	133.53%	88.06%	101.95%	127.98%	95.01%	118.56%
lane 7		Left	42.52%	38.10%	55.90%	36.07%	35.07%	46.65%	38.12%	41.53%	42.39%	41.78%
lane 8		Left	44.14%	55.23%	33.43%	50.44%	61.89%	40.43%	35.13%	49.03%	37.78%	45.81%
lane 9	South	Throughput	32.31%	65.90%	53.34%	64.66%	31.37%	40.70%	27.32%	49.52%	34.01%	45.09%
lane 10	stream	Throughput	28.50%	33.15%	27.90%	44.84%	23.81%	27.26%	32.21%	31.64%	29.74%	31.10%
lane 11		Throughput	29.09%	33.98%	34.78%	25.27%	34.80%	29.61%	24.37%	31.58%	26.99%	30.27%
lane 12		Right	140.04%	99.02%	104.30%	79.84%	87.94%	95.24%	70.66%	102.23%	82.95%	96.72%
lane 13		Left	79.54%	69.81%	81.42%	68.91%	62.29%	52.81%	61.42%	72.39%	57.12%	68.03%
lane 14		Left	69.34%	59.28%	62.76%	61.58%	83.38%	53.62%	58.56%	67.27%	56.09%	64.07%
lane 15	East	Throughput	97.17%	83.53%	79.72%	61.74%	72.62%	65.44%	55.77%	78.96%	60.61%	73.71%
lane 16	stream	Throughput	90.51%	69.15%	82.77%	71.05%	83.76%	62.56%	55.49%	79.45%	59.03%	73.61%
lane 17		Throughput	47.15%	51.87%	53.73%	61.65%	57.22%	54.34%	66.47%	54.32%	60.41%	56.06%
lane 18		Right	104.60%	95.47%	113.93%	110.93%	90.63%	95.06%	100.22%	103.11%	97.64%	101.55%
lane 19		Left	120.49%	128.70%	119.89%	131.11%	118.02%	76.65%	100.54%	123.64%	88.60%	113.63%
lane 20		Left	83.12%	127.65%	116.90%	116.21%	94.85%	113.27%	78.31%	107.75%	95.79%	104.33%
lane 21	North	Throughput	302.09%	65.40%	59.59%	79.06%	50.39%	47.24%	52.19%	111.31%	49.72%	93.71%
lane 22	stream	Throughput	64.75%	64.91%	86.81%	73.69%	61.91%	66.78%	69.24%	70.41%	68.01%	69.73%
lane 23		Throughput	65.88%	66.63%	67.62%	37.27%	66.49%	45.33%	41.27%	60.78%	43.30%	55.78%
lane 24		Right	63.88%	52.08%	51.52%	69.25%	42.54%	53.71%	45.51%	55.85%	49.61%	54.07%
AVG Left			66.53%	73.43%	72.84%	70.45%	71.42%	61.03%	59.62%	70.93%	60.33%	67.90%
AVG Throughput			79.38%	60.27%	61.47%	58.34%	55.35%	52.21%	49.50%	62.96%	50.85%	59.50%
AVG Right			105.01%	97.32%	102.91%	92.57%	88.66%	83.02%	79.59%	97.29%	81.30%	92.72%
AVG West			66.61%	73.35%	74.28%	65.07%	71.86%	63.35%	62.75%	70.23%	63.05%	68.18%
AVG South			52.77%	54.23%	51.61%	50.19%	45.81%	46.65%	37.97%	50.92%	42.31%	48.46%
AVG East			81.39%	71.52%	79.06%	72.64%	74.98%	63.97%	66.32%	75.92%	65.15%	72.84%
AVG North			116.70%	84.23%	83.72%	84.43%	72.37%	67.16%	64.51%	88.29%	65.84%	81.87%
AVG Total lanes			79.37%	70.83%	72.17%	68.08%	66.25%	60.28%	57.89%	71.34%	59.09%	67.84%

Appendix 5 Iteration

Resolution of 15/30 minutes in long term/short term

- ***Long-term***

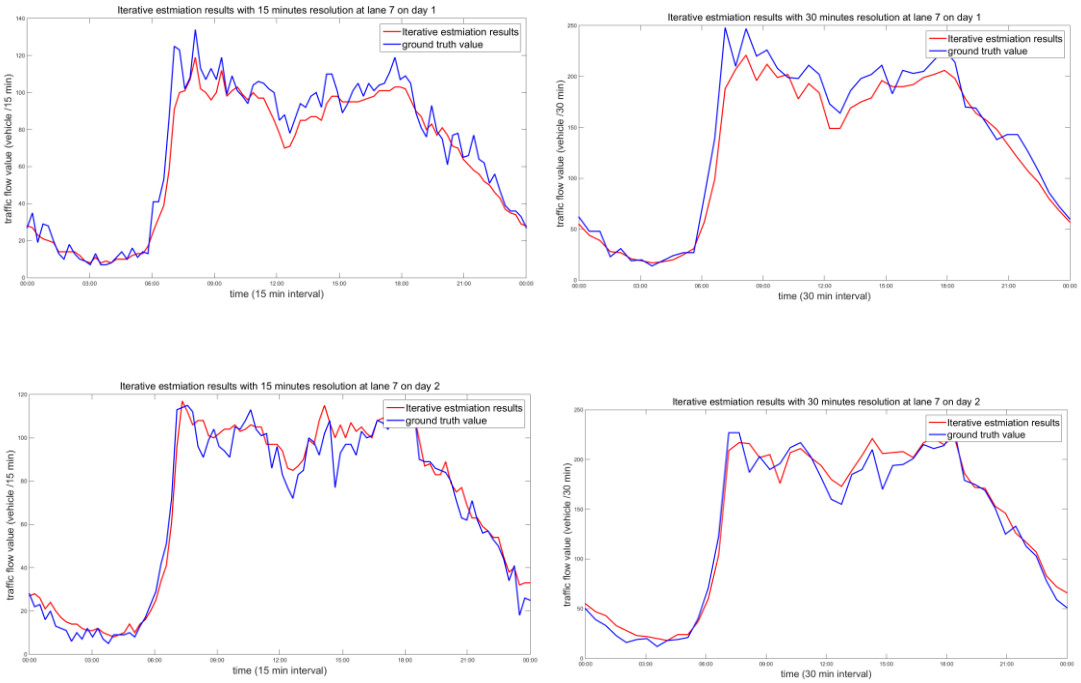


Figure 0-5 iterative estimation for long-term missing, on lane 7, day 15th and 23rd April 2013 ,15 minutes resolution (left)and 30 minutes resolution (right)

Table 0-19 error indicators for long-term missing, on lane 7, day 15th and 23rd April 2013, 5 15 30minutes resolution,

	5 min			**15 min**			**30 min**		
Error indicator	A1	A 2	Iterative results	A1	A2	Iterative results	A1	A 2	Iterative results
MAPE	29.97%	24.83%	25.16%	11.38%	13.53%	11.34%	10.94%	12.61%	10.15%
RMSE	6.66	4.95	5.18	9.65	10.53	9.17	30.43	19.55	16.89
MAPE	34.70%	27.12%	27.52%	16.03%	15.70%	14.37%	14.83%	13.28%	12.36%
RMSE	6.90	5.03	4.93	7.06	8.45	6.95	20.12	14.62	12.07

- ***Short-term***

Morning peak 7:00-10:00

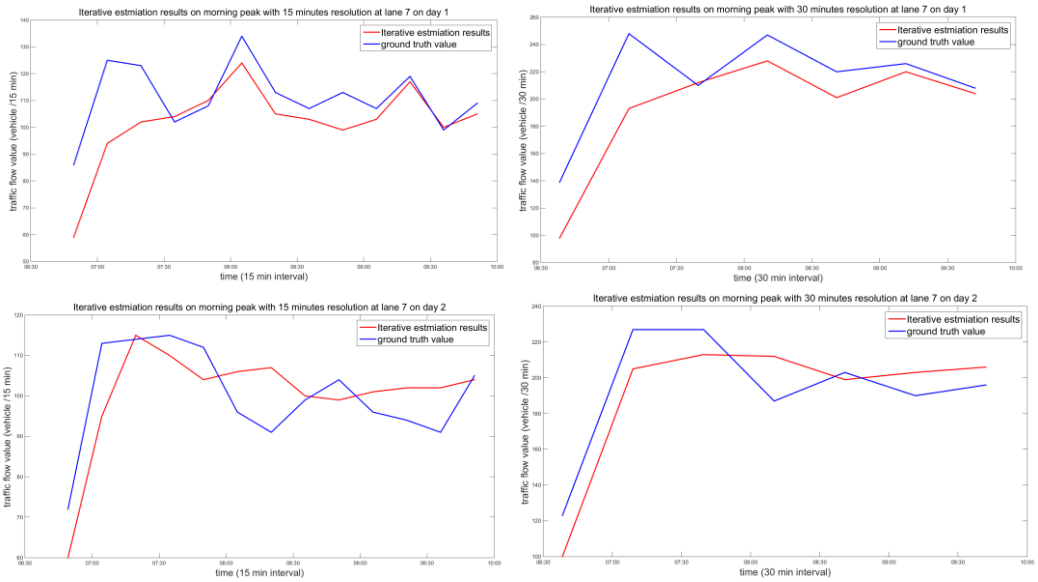


Figure 0-6 iterative estimation for short-term missing morning peak 7:00-10:00, on lane 7, day 15th and 23rd April 2013 ,15 minutes resolution (left)and 30 minutes resolution (right)

Afternoon peak 16:00-19:00

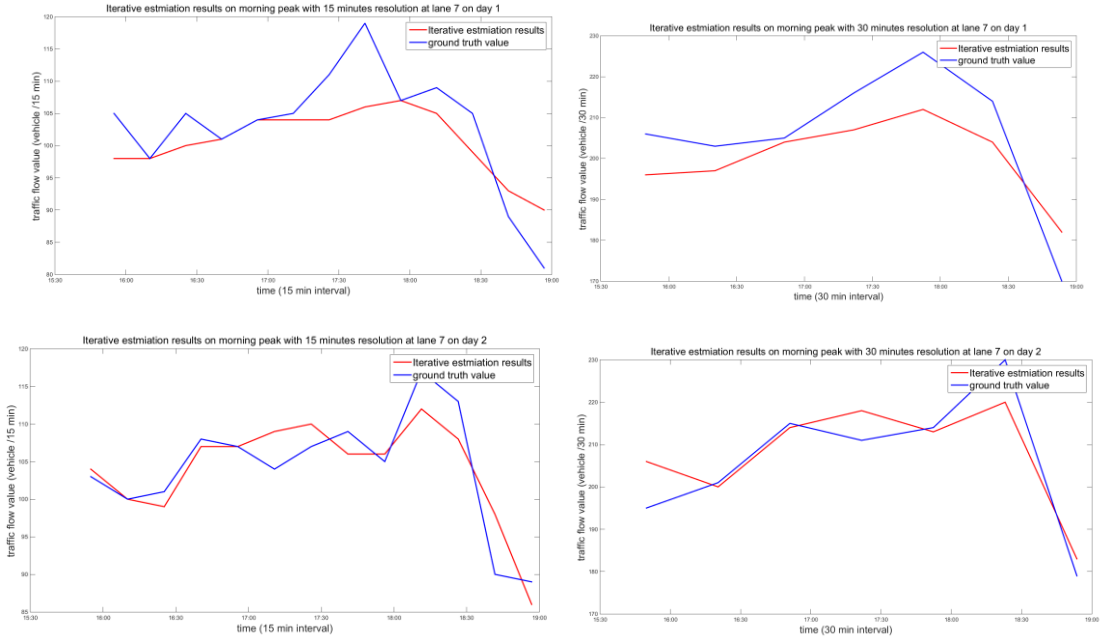


Figure 0-7 iterative estimation for short-term missing afternoon peak 16:00-19:00, on lane 7, day 15th and 23rd April 2013 ,15 minutes resolution (left)and 30 minutes resolution (right)

Appendix 6 Correlation coefficient map of traffic flow

Traffic flow over time

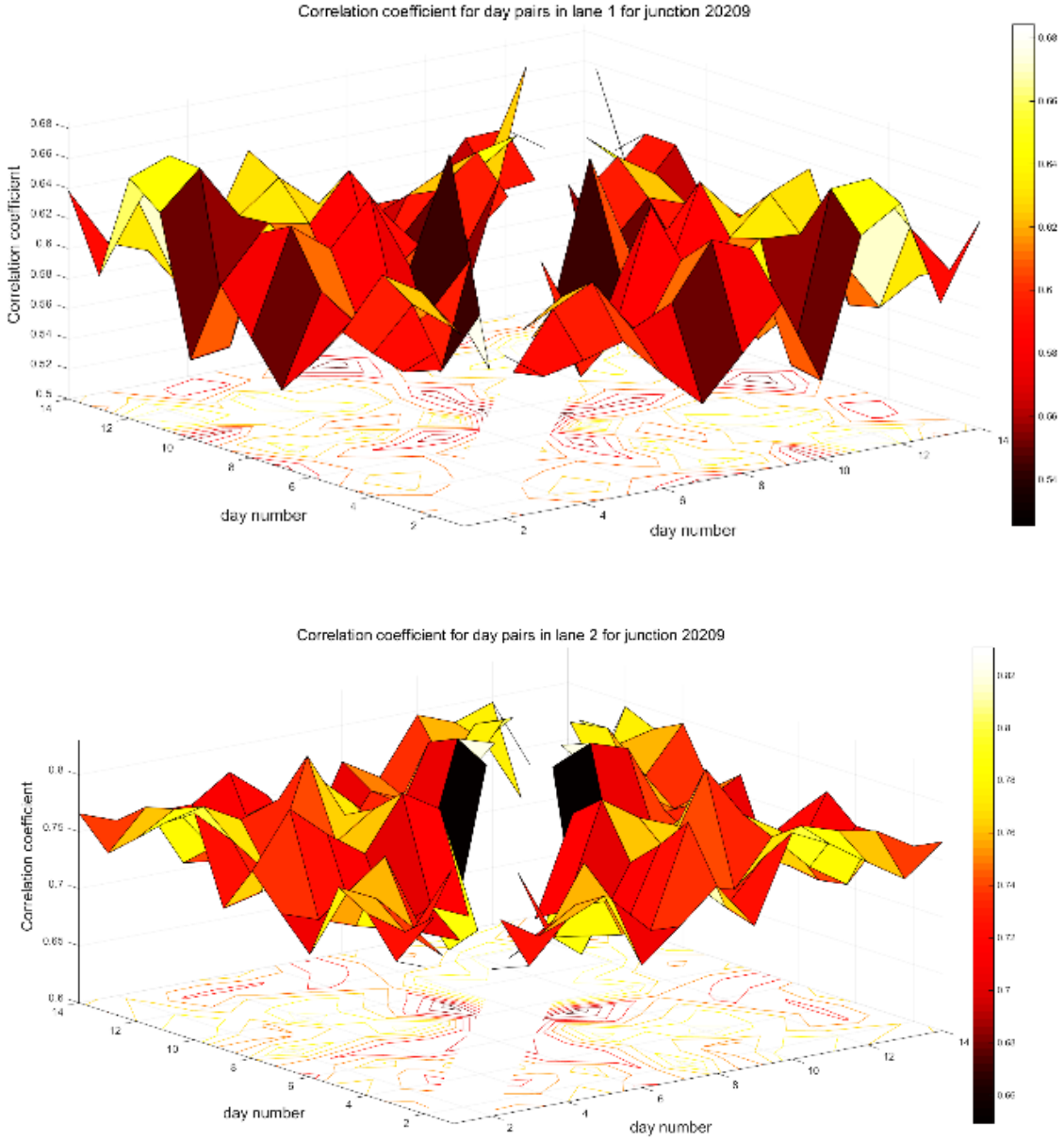


Figure 0-8 Correlation coefficient map for lane 1 and lane 2 at junction 20209

Traffic flow over spatial

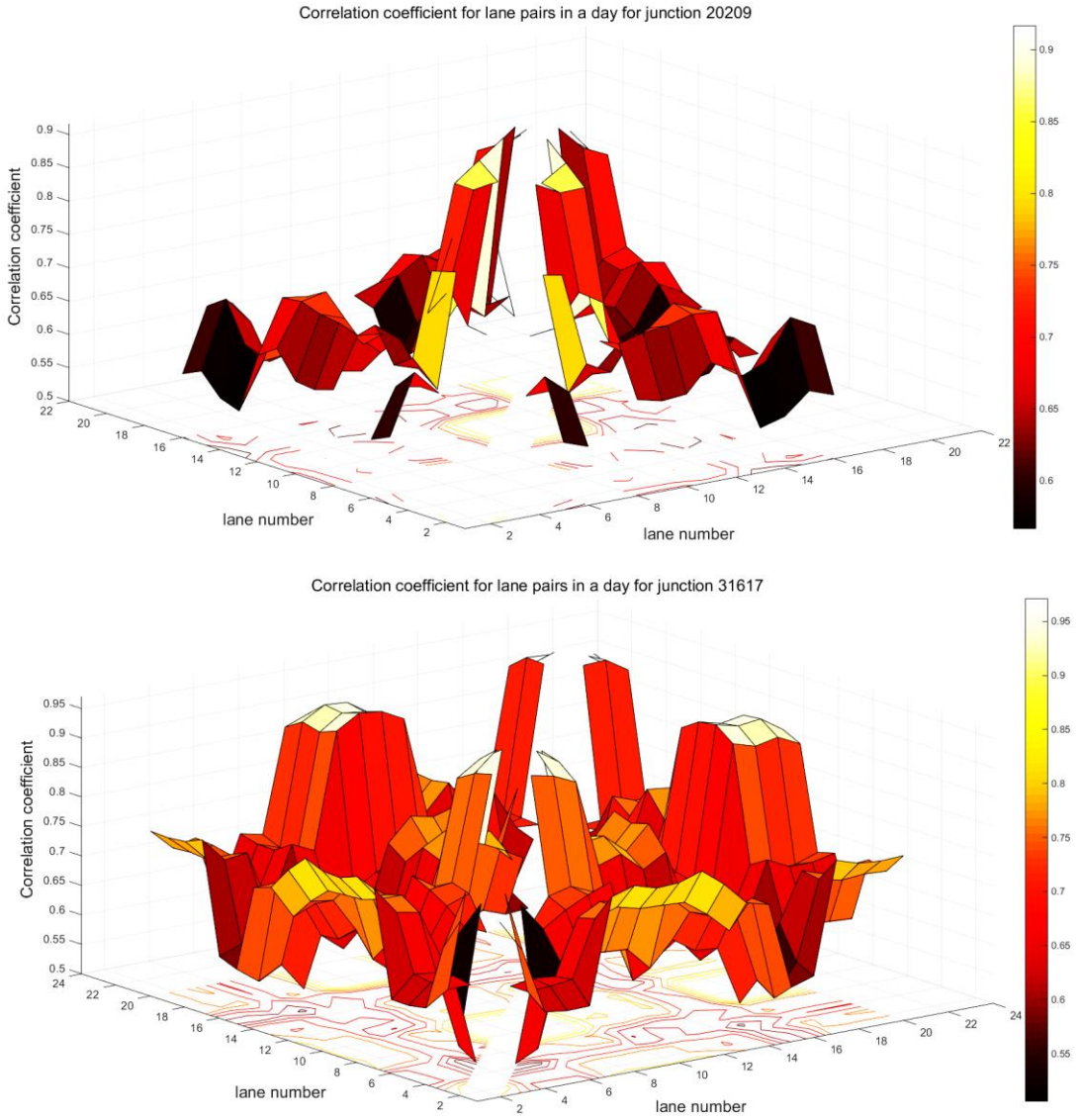


Figure 0-9 Correlation coefficient map for junction 20209 and 31617