# A Context-Aware Traffic Congestion Estimation Framework to Overcome Missing Sensory Data in Bangkok

by

**Panraphee Raphiphan**

BCompSc, MCompSc

Submitted in fulfilment of the requirements for the degree of

**Doctor of Philosophy**

June 2015

**Caulfield School of Information Technology**

**Monash University**

# Table of Contents

# Abstract

Traffic congestion is a problematic experience for commuters in metropolitan areas, costing time and money. For situations involving emergency services, traffic congestion may also be life threatening. A traffic information system (TIS) can play a significant role in improving traffic congestion problems by providing information to road users about the location and degree of traffic congestion. Knowing the location and degree of traffic congestion in real time can assist drivers in planning their routes to avoid heavy traffic. A TIS collects data from multiple sources, including sensors. However, data from sensors may become unavailable due to some reasons such as sensor damage or lost communication. In addition, some roads lack sensors. To ensure the availability and continuity of reported traffic information despite the uncertainty of sensor data, an approach that estimates traffic congestion when sensory data is not available is required.

In this thesis, we conduct research into this issue through the lens of a design science research methodology. We propose an artefact, the Context-Aware Traffic Congestion Estimation Framework to Overcome Missing Sensory Data (the CATE framework), to address the above issues. Most existing methods estimate traffic congestion using sensors. In contrast, the CATE framework utilizes available external context information to infer the traffic situation. The framework contains several inference models that represent different situations based on the available context. When sensory traffic data is missing, an appropriate model is selected during run time to infer the traffic congestion degree. The models were developed using machine learning algorithms during our research based on traffic data collected in Bangkok. To deal with the possibility of changes to traffic situations that may make predictions less accurate, the CATE framework incorporates a built-in relearning function that can be used to improve the accuracy of models over time.

During the test phase of this research, the CATE framework proved feasible and efficient. It inferred the traffic congestion degree with accuracy higher than that of existing methods and within comparable turnaround times.

To further improve the initial artefact of the CATE framework, further test was carried out in the form of survey. The survey aims to validate and improve the initial selection of the context information chosen for the CATE framework. The survey collected Bangkok road users' perceptions of the factors that affect traffic in Bangkok. The evaluation of this phase demonstrated that the final artefact improved from the initial artefact and again performed better than existing methods in terms of accuracy while also reducing the required processing times and costs associated with calculating the traffic congestion degree.

The proliferation of social media and mobile devices suggests that these are possible outlets for disseminating traffic reports in the future and so we included questions in our survey to investigate this possibility. We used the results of these questions to create recommendations for the development of TIS and traffic report services. These recommendations – that information regarding journey routes and traffic conditions be accessible via mobile devices and websites to meet the needs of road users, and that social networks be considered alternative sources of potential traffic data – can be used as guidelines to improve existing TIS and traffic information dissemination services in Bangkok.

Through the conceptualization and evaluation of our CATE framework, this thesis makes theoretical and practical contributions to the Intelligent Transportation System (ITS) domain. Through the survey based on the perceptions of Bangkok road users and subsequent statistical analysis, the thesis also makes contributions to the development of TIS reporting systems. Although our study was based on Bangkok data, it may be applicable to other cities that share similar road infrastructure and traffic information issues.

This research has produced a framework that has the potential to make a positive difference to road users. The results justify continuing research in this area in order to

increase the body of scientific knowledge of the ITS domain and to provide practical support to those involved in managing and maintaining TIS.

**Keywords:** Intelligent Transportation System, ITS, Traffic Information System, TIS, Traffic Report, Missing Data, Pervasive Computing, Context-Aware, Traffic Congestion, Intelligent System, Bangkok

# Declaration

This thesis contains no material which has been accepted for the award of any other degree or diploma in any university or other institution and affirms that to the best of my knowledge, the thesis contains no material previously published or written by another person, except where due reference is made in the text of thesis.

Panraphee Raphiphan

Date: __/___/___

# Acknowledgements

First I would like to express my sincere gratitude to my supervisors Dr Maria Indrawan-Santiago and Professor Arkady Zaslavsky for their guidance and support throughout these challenging years. Their valuable time and efforts are highly appreciated.

I also give my sincere thanks to Dr Sucha Smanchat, Dr Passakon Prathombutr, Associate Professor Phayung Meesad, Dr Wasan Pattara-atikom, Dr Anyarat Boonnithivorakul and Dr Rattanan Nantiyakul for providing useful advice and assistance. My thanks also go to the National Electronics and Computer Technology Center (NECTEC) and the Thai Meteorological Department for supplying the real data for my experiments, and to the respondents who participated in the online survey.

I am also grateful to Dr Megan Seen who proofread and edited this thesis. I also thank the staff at the Caulfield School of Information Technology for their support, and especially Ms Allison Mitchell who always gave me useful information and much-appreciated assistance.

An enormous thank you to all my friends who understand me and supported me through this challenging period. Last but not least, very special appreciation to my mom, my dad and my husband who have always been there to cheer me up and support me. Without you, I could not have come this far.

# List of Publications

Panraphee Raphiphan, Maria Indrawan-Santiago, Sucha Smanchat, and Arkady Zaslavsky, "CATE: A Framework for Traffic Congestion Estimation to Overcome Missing Sensory Data," *Intelligent Transport Systems, IET*. (Under Review)

Panraphee Raphiphan, Arkady Zaslavsky, and Maria Indrawan-Santiago, "Building Knowledge from Social Networks on What is Important to Drivers in Constrained Road Infrastructure," *Proc. of 18th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems*, Gdynia, Poland, September 2014.

Panraphee Raphiphan, Passakon Prathombutr, Arkady Zaslavsky, and Phayung Meesad, "Real Time Traffic Congestion Degree Computation for Minor Sensorless Roads Using Cost Efficient Context Reasoning," *Proc. of 13th International IEEE Conference on Intelligent Transportation Systems*, Madeira Island, Portugal, September 2010.

Panraphee Raphiphan, Arkady Zaslavsky, Passakon Prathombutr, and Phayung Meesad, "Overcoming Uncertainty of Roadside Sensors with Smart Adaptive Traffic Congestion Analysis System," *Proc. of IEEE Intelligent Vehicles Symposium (IV'09)*, China, June 2009.

Panraphee Raphiphan, Arkady Zaslavsky, Passakon Prathombutr, and Phayung Meesad, "Context Aware Traffic Congestion Estimation to Compensate Intermittently Avaialble Mobile Sensors," *Proc. of International Workshop on Mobile Urban Sensing (MobiUS)*, Taiwan, May 2009.

Panraphee Raphiphan, Arkady Zaslavsky, Wasan Pattara-Atikom, and Passakon Prathombutr, "Distributed Context Aware System for Travel Time Estimation Based on Cellular Probes in Metropolitan Environment," *Proc. of the Fourth International Conference of Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON2007)*, May 2007.

Panraphee Raphiphan, Wasan Pattara-Atikom, and Passakon Prathombutr, "A Survey of Travel Time Estimation Techniques Based on Cellular Probes," *Proc. of NSTDA Annual Conference 2007 (NAC2007)*, NSTDA publishing, Pathumthani, March 2007.

# List of Figures

# List of Tables

# Chapter 1 Introduction

This thesis introduces a framework that increases the tolerance levels of traffic information systems (TIS) to uncertain situations. The research was motivated by the need for reliable traffic information. Most TIS currently rely on data from traffic sensors. However, data from sensors may be intermittent or even absent, due to causes such as damaged components, poor weather conditions or insufficient sensors. An alternative is to draw data from contexts available in the environment. This contextual data compensates for missing sensory traffic data and enables continuity of the traffic information reported to users. To create this contextual data and alleviate the limitations of existing traffic data, in this thesis we propose an artefact, the Context-Aware Traffic Congestion Estimation Framework to Overcome Missing Sensory Data (the CATE framework).

This chapter begins with an introduction to the problem domain and the motivation behind the thesis. Research questions and objectives are presented. The chapter then briefly describes the research approach and outlines the scope of the thesis. The significance, contributions and limitations are discussed. The chapter closes with an outline of the thesis structure.

## *1.1 Problem Domain and Motivation*

### 1.1.1 Background

Traffic congestion is a serious problem for commuters in metropolitan areas. In countries with constrained road infrastructure such as Thailand, Indonesia and the Philippines, the problem is exacerbated. Despite the increasing road infrastructure in these countries, the disproportionate growth of the number of vehicles compared to the available road infrastructure results in severe traffic congestion in metropolitan areas. In addition to the tangible fuel energy costs, the delays caused by traffic congestion result in losses in many sectors of the economy.

Bangkok, the capital city of Thailand, is one of the cities where traffic congestion is a critical problem. The traffic congestion in Thailand has been a national problem for four decades [1]. In 2012, Bangkok was voted the most grid-locked city in the world [2]. Bangkok is an old city where high density residential areas are combined with ineffective road networks. For much of its history, the city has grown without proper city planning. These factors, combined with a lack of efficient traffic management and poor driving discipline, result in Bangkok constantly suffering from heavy traffic congestion, especially during rush hours. This congestion has both tangible and intangible negative impacts on the economy and the quality of life and environment around Bangkok. Nachaiwieng [3] concluded that the total economic loss from jammed traffic in Bangkok was approximately THB 165,400 million/year.

Over the years, a number of authorities have been involved in addressing traffic issues in Bangkok. These include the Office of the National Economics and Social Development Board, the Office of the Commission for the Management of Road Traffic (OCMLT), the Office of the Civil Service Commission (OCSC), the Mass Rapid Transit Authority of Thailand (MRTA), the Transport Co. Ltd., the Bangkok Metropolitan Administration, the Ministry of Education, the State Railway of Thailand (SRT), the Department of Highway, the Marine Department, the Department of Land Transport, the Department of City Planning, the Expressway Authority of Thailand (EXAT), the Department of Public Works and Town and Country Planning, the Royal Thai Police, the Department of Provincial Administration and the Bangkok Mass Transit Authority (BMTA) [4]. Despite the efforts of these bodies, traffic issues remain problematic in Bangkok.

### 1.1.2 Intelligent Transportation Systems

To help manage traffic issues, an Intelligent Transportation System (ITS) which rely on information and communication technologies in mitigating traffic congestion, enhancing safety, and improving quality of environment [5] have been heavily studied. The Intelligent Transportation Systems (ITS) integrates application of communications, control and information processing technologies to the transportation system. ITS covers all elements of the transportation system (vehicle, the infrastructure, and the driver or users) interacting together dynamically [6]. ITS

involves a large number of research areas spread over technological sectors such as electronics, control, communications, sensing, robotics, signal processing and information systems [7].

The number and type of modules of an ITS will differ depending on the needs of the country and its cities. However, in [6] state that the common core of modules usually consists of Advanced Traffic Management Systems (ATMS), Advanced Traveller Information Systems (ATIS), Advanced Vehicle Control System (AVCS), Commercial Vehicle Operations (CVO), Advanced Public Transportation Systems (APTS) and Electronic Payment Systems (EPS). The ATIS applies real-time data to assist drivers in choosing and scheduling their traveling route and mode, and is the module of greatest interest for this thesis cite.

### 1.1.3 Traffic Information Systems

A Traffic Information System (TIS) can be considered a subset of an ITS. A TIS processes traffic-related data into understandable information and provides this information to other systems or users. In general, the main behaviours of a TIS can be defined as collecting traffic data from sensors (or any kind of available sources such as GPS and smartphones), processing or interpreting data and then disseminating information to other systems or users [8, 9]. A depiction of a conventional TIS can be found in Figure 2-3. The data from sensors (both stationary and mobile) is typically transferred to a central traffic information centre (TIC) where the current traffic situation is analysed. Once analysed, the results are then disseminated to users [9]. Traffic information can be disseminated through many methods including radio, electronic boards installed along the roads, websites or mobile applications.

### 1.1.4 Data Generation and Sensors

In order to generate data for a TIS, various types of sensors are employed. Sensors can be categorized into stationary sensors, mobile sensors and mixed sensors [10]. While all can collect data for a TIS, they each have their limitations.

Stationary sensors (such as inductive loops, ultrasonic sensors and pneumatic road tubes) are fixed, non-moving sensors. Consequently, they are generally inflexible and require extra equipment for installation [11]. Disruption to traffic flow during

installation and installation costs are also limitations of some stationary sensors. Furthermore, equipment such as fixed video cameras that generate pictures to be used with image-processing algorithms for traffic analysis generally do not perform well in poor weather conditions.

Mobile sensors for traffic data collection are another solution for TIS [12-15]. Mobile sensors can take many forms such as Global Positioning System (GPS) [12, 16], mobile phones [17-19], and Automatic Vehicle Location (AVL) [20]. Mobile sensors can alleviate the limitations of stationary sensors, but come with their own issues. For example, a mobile sensor (such as a mobile phone) may move away from the relevant road segment. Data from mobile sensors may thus have only intermittent availability.

Stationary and mobile sensors can be combined to obtain better traffic data and improve coverage level. For example, data (such as occupancy and volume data averaged over a limited time period for a particular road) can come from both fixed detectors and from GPS-equipped probe vehicles [21]. An increased number of data sources enables the veracity of the data to be confirmed and leads to more accurate traffic data.

### 1.1.5 ITS in Thailand

In Thailand, ITS projects have not been widely implemented comparing to other countries. In 2010, in an attempt to reduce the traffic jam on the express ways by reducing the time required to pay the toll, EXAT started using the electronic toll-collection (ETC) system "Easy Pass". This system uses radio-frequency identification (RFID) technology [22]. Similar ETC systems have been used on tolled roads, bridges and tunnels in many countries such as Norway, The United States, Portugal, United Kingdom, Japan, South Korea, and Australia.

Another ITS project in Thailand is Bangkok's TIS. In this system, real-time traffic images from closed-circuit television (CCTV) are analysed in the computer system. The analysis results are then disseminated to commuters [23]. In 2011, 12,377 CCTV cameras had been installed across Bangkok including 1,892 CCTV cameras at street intersections, 1,873 CCTV cameras at government offices and hospitals and 4,975 CCTV cameras at schools and universities [24]. Despite these numbers, not all of the

main roads in Bangkok have had CCTV cameras installed. Additionally, a portion of these cameras will be inoperable at any one time due to weather, theft and/or poor maintenance, with a subsequent distortion of images and consequent loss of traffic data available for analysis.

In Bangkok, information is presented to travellers/drivers via the Variable Message Board (VMB) [25]. With this information, commuters in Bangkok can plan their routes to make their travel more efficient by avoiding congested roads.

### 1.1.6 Traffic Congestion Estimation Techniques Using Sensors

Different techniques and implementations of TIS provide information on a variety of issues such as road traffic estimation, travel time estimation, traffic reporting and vehicle to vehicle communication. Common methods to calculate current traffic congestion rely on data from sensors such as inductive loops, magnetic detectors and cameras to detect the volume of vehicles on the roads. The vehicle density is typically reported to traffic control centres where it is subject to traffic congestion estimation techniques.

Techniques using video cameras to detect traffic congestion work by applying methods based on time-spatial images [26, 27]. These techniques analyse still images from video and capture the differences between frames. With this information, the traffic situation can be identified, along a spectrum, as congested or flowing. This is the technique employed by Bangkok's TIS.

A mobile sensor approach is another possible technique to provide traffic data. Mobile phones have been researched in [28-30] as an alternative source of traffic data. Using this technique, the positions of road users' mobile phones are collected. The time taken for a mobile phone to move from one place to another is measured and mapped to obtain the density of the road's traffic. Probe vehicles equipped with GPS can also be used to collect traffic speed information. However, for road links with sparse probe vehicle data, the estimated mean can be inaccurate due to the low sample size.

RFID sensors can also be used to supply traffic data. The traffic monitoring system proposed in [31] attaches active RFID tags to probe vehicles and uses roadside wireless devices to collect signals from the RFID tags. As a probe vehicle with an

RFID tag passes the wireless readers, the average trip time and vehicle speed can be calculated. The congestion level is then measured based on the calculated speed with respect to pre-configured values. The proposed method also takes the average waiting time and site incidents into account, so the congestion level can be calculated more accurately.

### 1.1.7 Limitations of Existing Traffic Estimation Techniques

The techniques described so far rely heavily on sensory data to provide road traffic conditions. However, sensory data can be missing at particular times due to three possible scenarios. First, when using static sensors, the traffic data from some sensors may be lost due to unsuccessful data transmission or poor weather conditions, or the sensors may be broken or stolen. Second, if the system relies on mobile sensors such as GPS equipment or cellular phones as traffic probes, data from these mobile sensors may only be intermittently available as sensors move into and out of monitored roads. Third, in many countries (and especially in developing countries), investment in sensors and infrastructure is limited by capital budget. Traffic authorities normally choose not to invest in traffic sensors for less major roads due to the lower volume of traffic on these roads. Nonetheless, these roads may still be important to road users and for the estimation of traffic conditions. We define these categories of roads as "*resource constrained roads*". This concept will be further explained in section in Section 4.1 in Chapter 4.

In summary, most existing traffic estimation approaches rely on data from sensors. Based on the results from our literature review, other types of data sources for traffic congestion estimation rarely appear to have been considered. Even though some existing works have proposed methods to compensate for missing traffic data, most of these are based on the traffic data of observed road segments, which usually comes from the traffic sensors. Surrounding contexts (that is, data other than traffic data from observed road segments) are rarely taken into account. In addition, an attempt to estimate the traffic condition in sensorless road segments has not been made in most works. Nor, to the best of our knowledge, does Bangkok's TIS provide a solution to missing sensory traffic data or a solution to road segments lacking sensors. Overall, research into an alternative approach - using influential context attributes for traffic

congestion estimation - has rarely been considered. In addition, the study about suitable influential context attributes for traffic condition estimation was rarely been done especially in Bangkok's circumstance.

To introduce a solution for the above-mentioned issues, two research questions were posed, as explained in the next section.

## *1.2 Research Questions and Objective*

The aim of this research is to introduce an efficient framework for TIS that can continuously provide traffic information to users even when sensory traffic data is absent. The framework achieves this by utilizing environmental contexts to provide surrogate traffic information.

To address the aim of this research the following research questions were articulated:

***What information can be collected from other sources to approximate missing sensory data?***

***How would data collected from other sources be processed to approximate the missing sensory data?***

This research thus investigates contexts other than sensory traffic data that influence road traffic congestion and that can be used to efficiently and accurately approximate the missing sensory traffic data. Furthermore, this research aims to explore a method that is capable of compensating for the missing sensory traffic data by utilizing the information collected from other sources. As such, the main objective of this thesis is to propose a context-aware traffic congestion estimation framework to overcome missing sensory data. The proposed framework should be resilient to high missing data rates, efficient and give accurate results.

## 1.3 Research Approach

In order to answer the research questions effectively, the design science research paradigm was used to guide the investigation. This thesis has adopted the design science research cycle proposed by Hevner [32, 33] as its overarching framework and the design science research steps proposed by Vaishnavi and Kuechler [34, 35] as the guide for the research process. The research process comprised awareness of problem, suggestion, artefact development, evaluation and conclusion. Our research methodology is presented fully in Chapter 3 where we describe and explain our research questions, approach and process in detail.

## 1.4 Scope of Research

### 1.4.1 Artefact development

As discussed, ITSs comprise many facets of traffic management. In this research, we focused our efforts on solving an issue inherent in existing TIS (the sub-component of an ITS that is responsible for providing traffic information to drivers): that of unreliable sensor data and consequent unreliable traffic information. To address the issue of unreliable sensor data, we concentrated our efforts on developing a framework that can provide compensatory contextual traffic data when sensory traffic data is lacking. The CATE framework can efficiently and accurately approximate traffic information while requiring only reasonable resources and remaining resilient to high missing data rates. The framework ensures traffic report services are able to provide complete, reliable and accurate traffic congestion estimation.

The top row of the following diagram illustrates the inputs and outputs of a typical TIS. The items in the dotted frames outline the components and processes of the CATE framework while the following discussion provides further detail about the scope of this thesis.

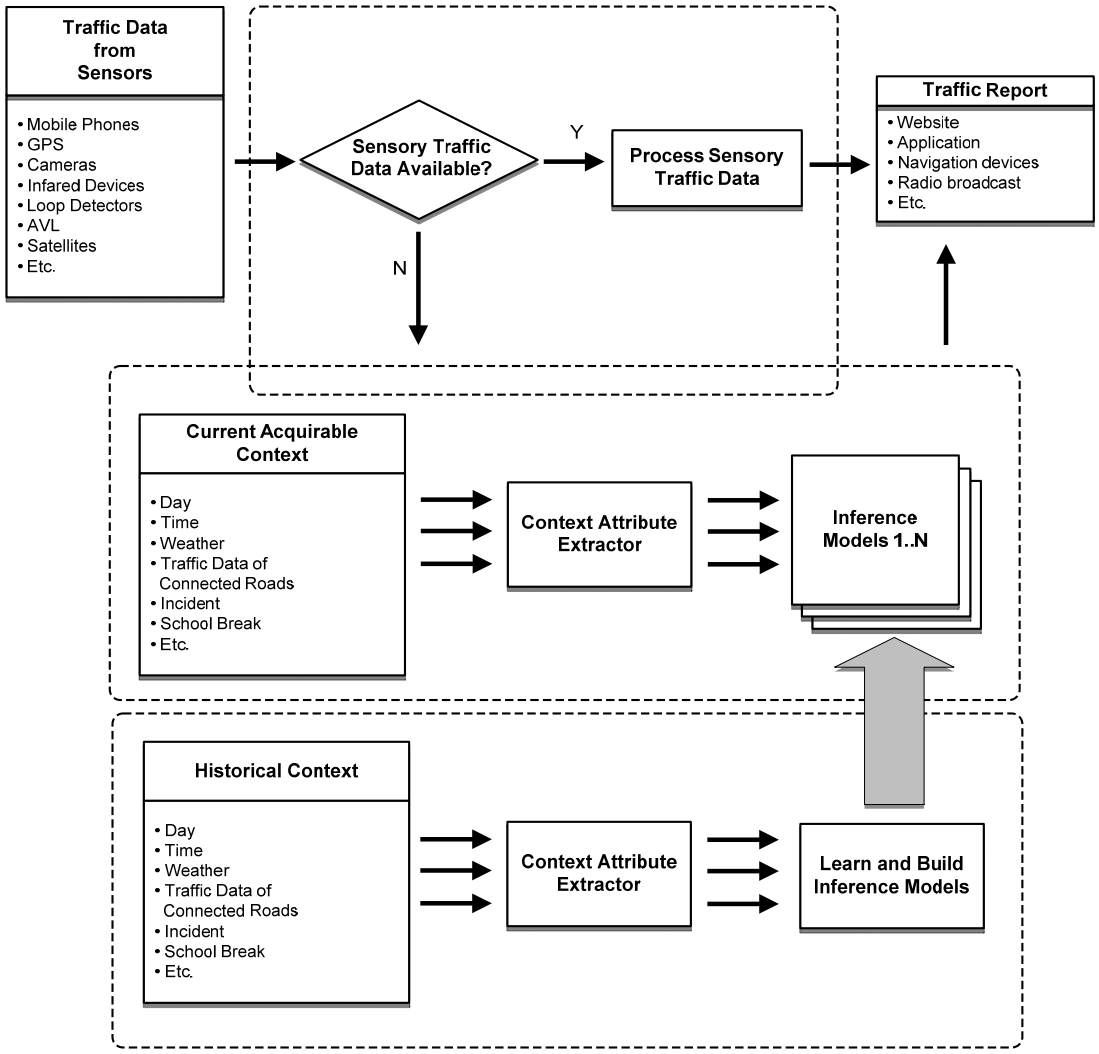



**Figure 1-1: Diagram of overall system to disseminate traffic information**

Figure 1-1 illustrates a system that provides traffic information to road users and illustrates its relationship to the CATE framework. The system collects data from the sensors in the street network (which may be roadside sensors, mobile sensors or a combination of both), processes the data and disseminates traffic information to road users. However, should sensory data be missing (due, for example to poor weather conditions), our framework approximates the missing sensory traffic data by using machine learning techniques along with contextual awareness. Thus, when sensory traffic data is missing, the current available context can be used to calculate the inferred traffic congestion degree. The context is any information that can be used to characterize the situation of an entity, ranging from low-level parameters to highly abstract concepts [36], [37]. A context attribute is any type of data that is used in the

process of calculation. The proposed framework comprises an *inference model building phase* and a *real time traffic inference phase*. The input for both phases is an influential context, which is a context that has impact on the monitored road.

Our framework starts by building different inference models from historical data. Each model is created for each set of current context attributes available at run-time to avoid accuracy degradation caused by missing values in a model. We also identified a reduced set of context attributes (*RS*) that acts as input to the traffic congestion inference process in order to improve our proposed framework.

When sensory traffic data is missing, a suitable inference model will be chosen in run-time depending on the available contexts at that time. Once a suitable model is selected, the current available context attributes are used as input for a selected inference model to generate an *inferred traffic congestion degree*. The inferred traffic congestion degree is then reported to users.

To deal with the possibility of changes to the situation and environment that affect traffic situations, the CATE framework incorporates a relearning process that can automatically adjust the inference models as the pool of actual data grows. Thus accuracy can be improved over time. Additionally, this technique allows the CATE framework to operate in situations even where historical data is sparse (such as road segments with newly installed traffic sensors). However, the optimal time period for relearning remains a topic for future research.

### 1.4.2 Identifying Traffic Information Needs and Alternative Methods of Traffic Data Collection and Dissemination

A TIS is concerned with both collecting input traffic data for analysis and disseminating consequent traffic information. In addition to developing the artefact to help inform the TIS to improve the quality of traffic information, the scope of our research also included identifying the traffic information needs of Bangkok's road users through an on-line survey. We also analysed and discussed guidelines for designing mobile applications and websites to disseminate traffic information and meet these needs. These analyses is not only useful for extending our proposed framework in the future, it is also beneficial for future implementations of traffic

report services for Bangkok's TIS. They may also help improve other existing or newly implemented TIS.

Additionally, this research also considered the feasibility of extracting and disseminating traffic data though social networks. This approach has been reported previously [38, 39], but not in the Bangkok context. Our study indicates that it is feasible to extract traffic data relevant to Bangkok from social networks [40]. It may be possible for future research to consider this issue in more detail by taking into account traffic information obtained from social network elements such as posts, comments and geographic locations through techniques such as text mining. However, this topic was not part of the current research.

## *1.5 Research Contributions*

The methods that we used and the outcomes of our research make contributions to both ITS knowledge and the general community in many aspects. We expound the contribution and significance of each, as follows.

### 1.5.1 Contribution to ITS Knowledge

This research provides a better understanding of the use of design science as a research methodology for solving problems in the ITS domain. It shows that design science can address and solve ongoing theoretical issues in ITS. Thus, it contributes to the theory of doing research in ITS.

Furthermore, this research has contributed novel findings for TIS within the ITS community. Our CATE framework can solve the problem of missing sensory traffic data in TIS. Our framework can be applied to existing TIS to improve the continuity of traffic information reported to users even when traffic sensors are unable to provide traffic data. The proposed framework also contributes to the knowledge of ITS, especially under circumstances of constrained resources, as our finding has added the new knowledge to the ITS knowledge based.

In addition, our proposed reduced set of influential context attributes (*RS*) not only improved our CATE framework to yield a refined framework, but selected *RS* also contribute to the domain knowledge of ITS by defining a minimum sufficient set of

factors required to estimate traffic conditions. In addition, anyone who desires to implement a TIS can use proposed *RS* as the guideline on which context should be collected as input for traffic congestion estimation. Likewise, the method we used to obtain the *RS* is also useful to traffic management authorities for analysing the influential context attributes to identify which are worth to invest on to obtain them.

Our research also involved statistical analysis of data from our online survey. The analysis focused on the traffic information needs of Bangkok's road users. The results from the analysis provide useful guideline for future implementations of TIS. We also analysed and discussed guidelines for designing mobile applications and websites to provide traffic information. These analyses are useful for future implementations of traffic report services for Bangkok's TIS. They also help improve other existing or newly implemented TIS. These are contributions to ITS knowledge.

Finally, the findings from our study on utilizing the data Bangkok road users post on social networks for TIS also contribute to knowledge of the TIS implementation domain. It offers some insights on how we can utilize data on social network for TIS.

### 1.5.2 Contribution to the General Community

Our survey results presented in Section 6.2 in Chapter 6 indicate that over 94% of road users recognize the usefulness of knowing traffic information. The outcome of the proposed framework and this study can help improve both the quality of traffic information system and its dissemination. In cities where traffic congestion is recognised as a major problem, knowing traffic conditions benefits not only road users, but also the economy, the environment and quality of life for city residents. Users can make better decisions when route planning to avoid congested roads, thus yielding time-saving and fuel-cost benefits. In some situations, such as those involving emergency services, knowing more complete traffic information could save lives. In addition, traffic management authorities, no longer constrained by limited numbers of road sensors, can supply traffic information for a greater range of roads, and can have greater confidence in the information being provided to road users.

## 1.6 Limitations and Applicability

Limitations of this research are mainly about data. The data we used in order to build the model was drawn from Bangkok traffic in 2009. The framework may have less relevance to TIS used in cities dissimilar to Bangkok. Conversely, the CATE framework may have greater applicability to TIS implemented in cities that face similar road infrastructure issues as Bangkok.

Another limitation relates to the data collected though our online survey. We used technology to collect data for our online survey, and thus ecludes the portion of the community not comfortable with using technology. In addition, question responses reflect the perceptions of respondents and consequently incorporate respondents' own bias and world view. While we have used these responses to inform guidelines for TIS implementation, further research to confirm our findings would be suggested.

## 1.7 Thesis Outline

This thesis comprises eight chapters. The structure of this thesis is as follows:

- Chapter 2

  This chapter reviews relevant literature and prior research. Knowledge relating to designing our artefact is presented in this chapter including discussion. The limitations of existing works are also observed and specified here.

- Chapter 3

  This chapter discusses the research methodology adopted for the current research. The research questions are presented. This is followed by a discussion of design science as a research method. Finally, the design science research processes including awareness of problem, suggestion, artefact development, evaluation and conclusion are explained.

- Chapter 4

  This chapter describes our designed artefact, the Context-Aware Traffic Congestion Estimation Framework to Overcome Missing Sensory Data (the CATE framework). An analysis of suitable machine learning algorithms for designing our

proposed framework is also presented. Other existing approaches to compensate for missing data are also discussed. These are later compared to our proposed framework in evaluation.

- Chapter 5

  Chapter 5 conducts the first evaluation (Evaluation I) of our proposed framework. The simulation of the real situation used for our evaluation is explained. The evaluation setup, evaluation method and the evaluation results comparing our approach with other approaches are reported. We evaluate our proposed framework in terms of its feasibility and efficiency compared to other approaches. The findings from the evaluation in this chapter lead to suggestions and conclusions that merit further investigation to improve the proposed framework.

- Chapter 6

  Chapter 6 investigates improving the proposed framework by analysing a reduced set of context attributes to be used for the framework. This process yields the final artefact. The refined CATE framework is then proposed, followed by an evaluation. The evaluation also compares the performance of the CATE framework against the performance of other approaches.

- Chapter 7

  In addition to designing the artefact, we performed further research to study the traffic information needs of users and their preferred communication channels. We also explored the potential use of social networks for TIS in Bangkok in order to facilitate implementation of an efficient TIS and traffic report services in the future.

- Chapter 8

  This last chapter summarizes how the research findings address the research questions. In the chapter, the theoretical, methodological and practical contributions of this research are also discussed. The adaptation of Hevner et al.'s seven design science research guidelines [33] to this research is elaborated to

illustrate the quality of this research. This chapter ends with suggestions relating to directions for future research.

## 1.8 Chapter Summary

This chapter summarizes the motivation for this research, the research questions and our research approach in solving the research questions extracted from the problem domain and motivation. The research approach was briefly presented in terms of how the research questions were explored within a design science research framework. Our research was motivated by the need to compensate for missing sensory traffic data and provide continuity of traffic information for users. Most existing approaches rely on sensory data. However, sensory data may not be available due to many uncontrollable reasons. We thus proposed the Context-Aware Traffic Congestion Estimation Framework to Overcome Missing Sensory Data (the CATE framework) which uses available external contexts in the environment instead of solely relying on sensory data.

This chapter also includes a description of the scope of the thesis. A summary of the research significance of this thesis and its contributions was briefly discussed and its limitations were described. Finally, the structure of the thesis, which illustrates the whole research process and the role of each chapter, was explained.

# Chapter 2 Related Works

In this chapter related works and background information are presented in order to describe the problem domain. The knowledge base used for designing our artefact is also included in this chapter. We begin by discussing intelligent transport systems (ITS) to lay out the background of the advanced applications that provide traffic management services and allow users to be better informed. Associated technologies and electronic devices (including stationary sensors, mobile sensors and mixed sensors that collect traffic data) are then described. The data collected from these sensors requires calculation, processing, interpretation and dissemination, which leads to a discussion of traffic information system (TIS). In Section 0, different traffic congestion estimation techniques are presented. Methods and concepts (pervasive computing, context awareness, data mining tasks and machine learning algorithms) related to design of our artefact are explained in the last section of this chapter.

## *2.1 Intelligent Transport Systems*

The consequences of excessive vehicles on roads range from the more trivial (such as personal inconvenience resulting from delays caused by heavy traffic) to the critical and important (such as accidents, emergency vehicles being unable to respond to life-threatening situations and pollution emission) [41].

ITS were first proposed in the 1930s as a possible approach to achieve the efficient utilization of transport infrastructure. The concept involved applying a radar system to automate a road vehicle to operate by itself [42], [43]. At that time, however, the concept did not gain much interest because the construction of new roads seemed to be a better approach to handling traffic congestion. ITS gained support in 1986 when governments, companies and universities from European countries established the Prometheus (Program for European Traffic with Highest Efficiency and Unprecedented Safety) project [42]. As a consequence, a prototype vehicle named VaMoRs [44] and the later VITA II [45] was introduced. Both prototypes were fitted with a number of cameras and processors so that they could operate in the middle of a

lane while a safe distance from the car in front and change lanes and overtake other vehicles.

Since the 1990s, ITS technology has involved roads and users as well as intelligent vehicles. The modern trend of ITS has considered large-scale integration and deployment [41] to offer efficient transport management systems in collaboration with advanced information technology, data transmission technology, electric sensor technology, electric control technology and computer processing technology [46].

Figure 2-1 illustrates the main idea of ITS where users, roads and vehicles collaborate and share traffic information for both tangible and intangible benefits including transport efficiency, comfort, safety and environmental conservation [41].

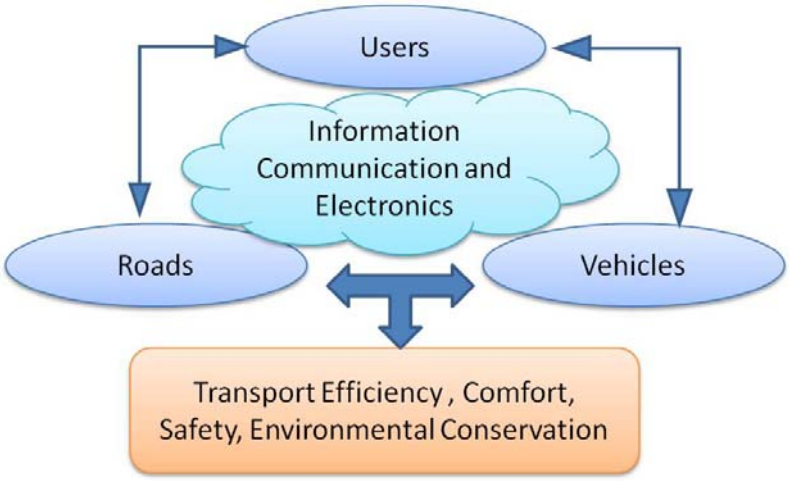



**Figure 2-1: ITS conceptual model [41]**

Type of modules of an ITS will differ depending on the needs of the country and its cities. In [41], ITS is categorized into 6 major groups as follows:

- **Advanced Traffic Management Systems (ATMS)**

  A purpose of ATMS is to provide real-time traffic information. Fundamental components of ATMS are series of video and roadway loop detectors, variable message signs, network signals and ramp meter timing schedules. All traffic data is processed from a control center involving real-time traffic control systems.

- **Advanced Traveller Information Systems (ATIS)**

  An ATIS provides traffic information to users to help them make better decisions about route selection. Examples of information include current traffic flows in specific areas, accidents and recommended shortest routes. The information can be displayed to users via the internet, radio, electronic devices equipped in vehicles or electronic boards installed along roads.

- **Commercial Vehicles Operation (CVO)**

  A CVO is an extension of real-time traffic information. Traffic data is used by logistic companies to manage and track their couriers and parcels.

- **Advanced Public Transportations Systems (APTS)**

  A combination of ATMS and ATIS, APTS result in better mass transport services. Transport can be appropriately managed and scheduled in response to circumstances in order to satisfy customers. For example, traffic lights may give priority to local buses or trains or local transportation services might be increased or decreased depending upon the needs of passengers at particular periods. Real-time locations of the buses can be viewed from boards at bus stops.

- **Advanced Vehicles Control Systems (AVCS)**

  The main purpose of AVCS is to achieve greater safety. However, when the number of accidents is reduced, traffic congestion resulting from accidents can also be reduced. Drivers can obtain traffic information from in-vehicle devices. In addition, vehicles also have computers to assist drivers such as cruise control, ABS and distance detectors.

- **Advanced Rural Transports Systems (ARTS)**

  Within rural areas, ATIS, ATMS and APTS are combined together to create ARTS to manage unique characteristics such as steep grades, blind corners, curves, few navigational signs, a mix of users and a lack of alternative routes.

All these systems require data to calculate traffic information. In order to obtain such data from the environment, various sensors can be used.

## 2.2 Sensor Technology in ITS

A sensor, in ITS terms, is an electronic device with associated software that detects vehicles and converts this data into traffic flow information. Sensors are a key component of ITS. Sensors act as traffic data collectors to provide data that underpin the whole system [47]. Sensors used for ITS can be categorized into stationary sensors, mobile sensors and mixed sensors [12-14, 16-21, 48], as summarized in Table 2-1.

**Table 2-1: Examples of sensor technologies in ITS**

| Stationary Sensors | Mobile Sensors | Mixed Sensors |
|---|---|---|
| Inductive Loops, Camera/ Video Image Processing, Microwave Radars, Passive and Active Infrared Devices, Ultrasonic Detectors, Passive Acoustic Detectors, Weigh-in-Motion Detectors, Magnetic Detectors, Pneumatic Road Tubes, Piezoelectric Detectors, etc. | Cellular Phones, GPS, Mobile devices equipped with video camera, Satellites, etc. | Loop Detector + GPS<br>Loop Detector + AVL, etc. |

### 2.2.1 Stationary Sensor Technologies in ITS

Various stationary sensors have been used in many countries for ITS. Nevertheless, most of these are inflexible and require the installation of extra equipment to existing infrastructures. Disruption to traffic flow and the time required for installation are also limitations of some stationary sensors. Furthermore, equipment such as fixed video cameras, which are incorporated with image processing algorithms to provide traffic flow information, do not perform properly in poor weather conditions.

We can categorize stationary sensors into intrusive detectors (which are installed in or under the road) and non-intrusive detectors (which are installed above or next to the road).

- Intrusive detector technologies include inductive loops, magnetic detectors, pneumatic road tubes, piezoelectric detectors and other weigh-in-motion (WIM) detectors. Intrusive detectors, such as inductive loops, have been widely used by

practical operators in past decades. Their limitations include disruption to traffic flow during installation and maintenance, high failure rates under certain conditions and inflexibility.

- Non-intrusive detector technologies include active and passive infrared, microwave radar, ultrasonic, passive acoustic and video image processing detectors. Active infrared, microwave radar and ultrasonic detectors are active detectors that transmit wave energy toward a target and measure the reflected wave. Passive infrared, passive acoustic and video image processing detectors are passive detectors that measure the energy emitted by a target or the image of the detection zone. When mounted at the side of a roadway, non-intrusive sensors cause less disruption than intrusive detectors, but installing sensors over a roadway requires lane closing for safety [11], [47].

### 2.2.2 Mobile Sensors Technologies in ITS

Using mobile sensors for traffic data collection is another approach for ITS [12-16]. To date, a probe vehicle and a remote sensing method have been most often described in studies. This section describes major technologies in probe vehicle approaches.

- Global Positioning System (GPS): A receiver uses signals transmitted from more than three satellites in order to determine the receiver's distance from each satellite. The distances are then calculated with a trilateration [19] technique to find the exact location of the receiver. Figure 2-2 shows a trilateration technique whereby the location of a mobile phone or GPS is calculated by distances from three satellite stations.

- Automatic Vehicle Location (AVL) on cellular phones: This technique uses a method similar GPS to determine location, but using cell sites instead of satellites [49].

- Automatic Vehicle Identification (AVI): A transponder device with its information is attached to a vehicle. Whenever a vehicle equipped with a transponder approaches the range of a roadside interrogator, information is automatically transferred from the transponder to the roadside interrogator, thereby uniquely identifying the vehicle [50].

- Remote Sensing: This technique collects data about objects or the landscape without direct physical contact. Traffic information can be collected via aircraft or satellites [11]. A remote sensing technology mostly relies on satellite images to extract traffic situations. Satellite images associated with image processing can provide traffic condition data. Prior research has focused on remote sensing technology to extract traffic conditions [11, 47, 51]. However, this technology faces the same problems as cameras that have limitations due to poor visibility in bad weather conditions such as fog or rain.

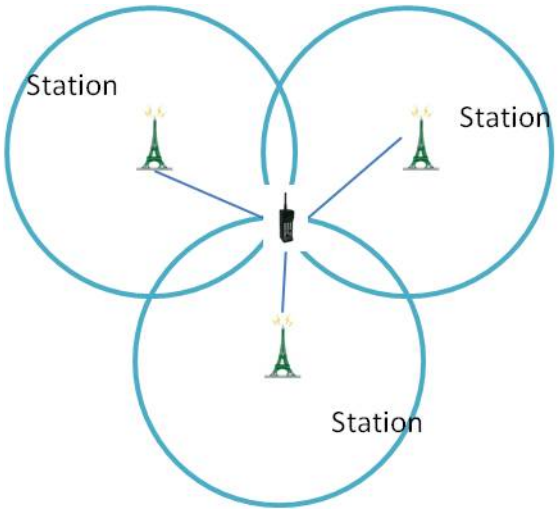



**Figure 2-2: Trilateration**

Mobile sensor techniques can alleviate limitations of stationary sensors such as inflexibility. However, relying on mobile sensors as the data provider comes with the problem of intermittent and hence unreliable availability. Due to its mobility, a sensor may move out of range. This problem can lead to an inability to disseminate traffic data of observed road segments at particular times due to the abruption of traffic data sources.

Techniques for using mobile sensors for traffic state estimation have been extensively studied. Zou, Xu, and Zhu developed a methodology to estimate speed and traveling time based on data from GPS receivers installed in taxis acting as probe vehicles in Guangzhou in China [12]. Cellular phones can also be used as mobile sensors to

estimate traffic information. Hsiao and Chang [52] introduced a segment based method using cellular network data to generate traffic speed information instead of a general distance based technique. Pattara-atikom, Peachavanish and Luckana [29] and Hongsakham Pattara-atikom and Peachavanish [30] proposed methods that use Cell Dwell Time (CDT) (the duration that a particular cellular phone remains in contact with a particular base station), a simple threshold or fuzzy logic technique and a neural network or K-means clustering technique respectively.

Even though promising results were obtained from the above works, they rely on the known sensory data of the road segment under observation (the road segment to be estimated) and generally do not involve other acquirable data in the environment. Nor do they attempt to estimate traffic conditions in sensorless road segments from other available known contexts.

### 2.2.3 Mixed Sensor Technologies

Stationary and mobile sensors can be used together to improve coverage level and obtain better traffic data. Collecting traffic data from more than one type of source thus has two advantages. Adding a mobile sensor to a system that already has fixed sensors enables the coverage to be expanded, and at a cost less than that of a fixed sensor, as shown in [53] and [20]. For example, data can come from fixed detectors (which provide occupancy and volume data averaged over a limited time period for specific sections of network links) while GPS-equipped probe vehicles (that is, additional mobile sensors) provide location data along with time data for adjacent road sections [21]. The second advantage is the improved accuracy of travel time estimation.

Mixed sensor technologies also have their drawbacks. When more than one type of sensor is used, data fusion algorithms are needed [21, 54]. Moreover, additional sensors of any type incur additional costs.

## *2.3 Traffic Information System (TIS)*

Raw traffic data gathered from sensors is of little use unless appropriately interpreted. Traffic information systems (TIS) were introduced to process traffic-related data into

understandable information and to provide this information to users or other systems. In general, the main behaviours of TIS can be defined as collecting traffic data from sensors (or any kind of available sources such as GPS and smartphones), processing or interpreting this data and then disseminating information to other systems or users. For instance, TIS can collect data from stationary sensors and vehicle-attached GPS and then combine and compute these two sets of data into a congestion level. This result is then shared with road users. A comprehensive TIS will process and share data in a number of areas, including transit routes, new roads, transit schedules, accidents, turning restrictions, incidents, traffic conditions, speed restrictions, operational, direction control statistics, lane closures, trends, road diversions, usage, delay time, congestion and travel time [8, 9].

Conventional TIS are organized in a centralistic way as illustrated in Figure 2-3. Some stationary sensors are deployed directly at the roadside. Mobile sensors in probe vehicles can also provide sensory traffic data. The data from sensors is transferred to a central Traffic Information Centre (TIC) and the current traffic situation is analysed. The result of this situation analysis is disseminated through channels such as radio broadcast stations or mobile phones [9].

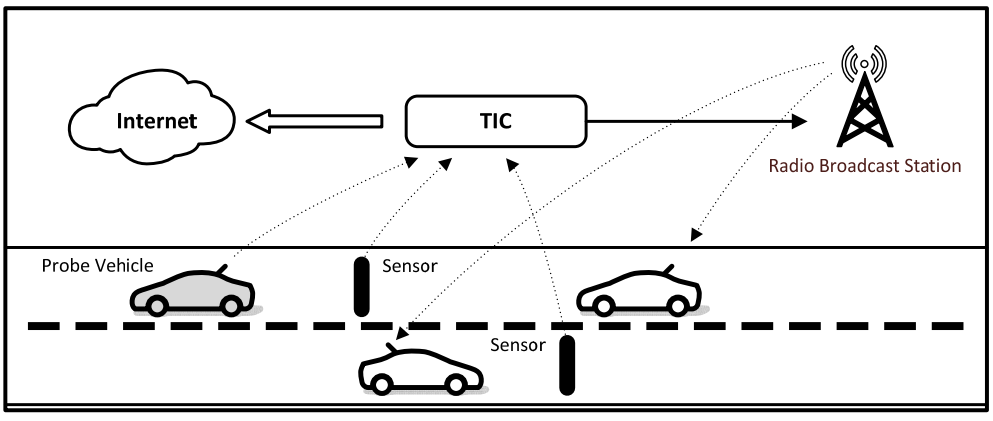



**Figure 2-3: Conventional centralized TIS [9]**

A number of related works have presented different techniques and implementations of TIS. This previous research has focused on various issues such as traffic congestion

estimation, travelling time estimation, traffic data dissemination and traffic reporting. The approaches relevant to this thesis are presented in the next section.

## 2.4 Road Traffic Estimation Approaches

Road traffic estimation approaches involve many issues such as traffic congestion estimation, travelling time estimation and traffic information dissemination (traffic reporting). Many definitions for traffic congestion have been proposed but a universal definition of traffic congestion has yet to be identified. Definitions of traffic congestion can be broadly categorized into three groups related to 1) demand capacity, 2) delays in travel time, and 3) costs [55]. An example of the first group is found in [56], where road traffic congestion is described as a condition in which the number of vehicles attempting to use a roadway at any time exceeds the ability of the roadway to carry the load at generally acceptable service levels. An example of the second group is found in Weisbrod, Vary and Treyz [57], who describe traffic congestion as traffic delay resulting from the number of vehicles trying to use the road exceeding the traffic network capacity. The phenomenon of increased auto travel time due to increased travel demand is also defined as traffic congestion in [58].

Over decades, many methods have been developed to collect real-time traffic data. Traffic data can be either quantitative (for example, lane velocity rates, queue lengths, trip times or waiting times) or qualitative (for example, incident descriptions). Common methods to estimate current traffic congestion rely on data from sensors such as inductive loops, magnetic detectors and cameras to detect the number of passing vehicles or the volume of vehicles on the roads. The vehicle density is reported to TICs before the current traffic situation is forwarded to road users via radio, GPS devices, internet web sites or mobile applications. In the past, ITS in Europe used inductive loops under the roads or magnetic detectors directly connected to traffic lights. The limitation of the system was that it did not provide information to other parts of the system. Modern techniques then have been introduced to solve this problem. Examples of these new approaches involve using video cameras to detect traffic congestion by applying methods based on time-spatial images [26, 27]. These images are sent to a TIC where the still images from video are analysed to capture

differences between frames so that the traffic situation can be identified as congested or flowing.

Techniques based on stationary sensors are considered powerful but can be impractical to install due to the associated expense and disturbance to commuters during installation. A mobile sensor approach enables traffic data to be obtained without the installation of sensors along roads.

Cellular phones have been researched in [28-30] as an alternative method to gather traffic data where no sensor is available. This works by collecting the position of road users' cellular phones and by measuring the time taken for the phones to move from place to place. The density of road traffic can then be obtained. In theory, the use of mobile phone data is feasible, but in some countries it is not acceptable due to privacy policies and sometimes the result may not accurately reflect the traffic situation (for example, if a commuter drives slower or faster than the rest of the traffic).

In contrast to mobile phones, purpose-installed GPS devices give more accurate results. With the reducing cost of technology, it is more practical to acquire data from GPS equipment rather than a cellular network.

Probe vehicle techniques are also proposed in many works. RFID can be another type of sensors. Sensors using RFID technology installed in probe vehicles have been described in Mandal et al. [31]. The Traffic Monitor system presented in this study uses active RFID tags kept in the probe vehicle and roadside wireless devices to collect signals from the active RFID tags attached to the probe vehicle. As the probe vehicle passes the wireless readers, the average trip time and vehicle speed can be calculated. The congestion level is then measured based on the calculated speed with respect to pre-configured values. The proposed method also takes the average waiting time and site incidents into account, so that the congestion level can be calculated more accurately. In 2010 Yanping, Dudu, Mingliang and Qi proposed a new method to identify the traffic congestion using high resolution satellite imagery [51]. Another approach proposed by Zhu and Li [59] focuses on reducing the calculation cost. The approach is to identify a moving object or probe vehicle using image processing techniques and then calculate the congestion based on its movement.

Speed estimation can also indicate the level of congestion. The higher the speed, the less the traffic congestion. Probe vehicles equipped with GPS can be used to collect traffic speed information and the statistical mean value of link speeds collected over time is often used as an estimator for mid-term predictions. However, for road links with sparse probe vehicle data, the estimated mean can be inaccurate due to the low sample size. Widhalm et al. have proposed a Gaussian-mixture based technique to increase the robustness of speed estimates in [60].

Bauza, Gozalvez and Sanchez-Soriano suggest that cooperative vehicles (that is, probe vehicles) could become valuable mobile sensors as the continuous exchange of messages between vehicles, including their locations and speed, can be a powerful tool. Probe vehicles can continuously monitor local road traffic conditions without requiring the deployment of a large number of infrastructure sensors or nodes. Bauza, Gozalvez and Sanchez-Soriano also proposed CoTEC (COperative Traffic congestion detECtion), a cooperative technique based on Vehicle-to-Vehicle (V2V) communication and fuzzy logic to detect road traffic congestion without deploying sensor infrastructure. CoTEC uses the beacon messages that vehicles periodically broadcast, mainly for safety purposes, to monitor road traffic conditions. Such techniques can accurately detect traffic congestion intensity and length [61].

A cloud theory based fuzzy identification method has been proposed by Wang, Xu and Zhang. This method constructs a traffic state judgment matrix based on a hierarchy of roads network under different congestion levels to analyse traffic congestion [62].

The research into TIS also includes travel time estimation techniques. Travel time can be calculated by the simple equation of *Time = Velocity / Distance*, but this formula cannot provide reliable time estimation due to fluctuating speeds resulting from traffic density. In reality, travelling time can be more accurately estimated with real-time traffic congestion information provided from various sensors.

The speed or velocity of vehicles can vary depending upon congestion levels, and thus the congestion level directly influences travel time. A macroscopic model used to calculate particular velocity is described in Kachroo, Ozbay and Hobeika [63] where a prediction-based travel-time function relies on free flow speed and congestion density.

The formula is written as *Approximate velocity = Velocity (free flow) * (1- current density/max density).*

Kachroo, Ozbay and Hobeika's study [63] also proposes two methods of traffic estimation using the infinite dimensional domain and space discretization (partial differential equation). Phan and Ferrie [64] and Sun et al. [65] have provided possible approaches to measure road density (or traffic flow state) by merging multisource traffic data (for example, historical traffic data, GIS databases, video detection and GPS based vehicles).

In 2010, an improved real-time travel time estimation system was launched in Stockholm [66]. This study gathered traffic information from an Automatic Vehicle Identification system (AVI) and processed this information with its own technique to provide better online travel time estimation with respect to real-time traffic information. In 2012 Tao et al. [67] introduced a technique using A-GPS in smartphones to continuously update the location of vehicles at regular intervals. This technique improved the accuracy of speed estimation for each road in a real-time approach. The tracking data obtained from individual mobile probes were aggregated to provide estimations of average road link speeds. The average speeds were then classified into different traffic condition levels to display on traffic reports.

van Lint in [68] proposed a method to predict travel time by using simulation and modelling. The calculation uses factors that are continuously present (for example, daily, weekly or yearly traffic patterns) and also event patterns (for example, public holidays, scheduled events (such as football matches or concerts), the level of daylight or snowfall rate). A modelling technique for travel time calculation was also proposed in Shange et al. [69]. A graph theory was applied where each vertex (or node) had its own travel time. The travel time from location A to location B could thus be calculated from the weights of the vertices existing on the path. In addition, the authors of this work also addressed the time delay between the vertices in their traffic-aware spatial network.

Those existing works provide a number of possible approaches for supporting TIS but they mostly rely on data from sensors and leave us still facing issues when sensor data is unavailable, whether due to factors such as poor weather or insufficient sensor

infrastructure. This can be an obstacle for producing and disseminating complete traffic information to users. In addition, when forecasting traffic, if traffic data is lacking, the ability to predict performance will fall sharply [70]. A means to estimate accurate and reliable traffic data in the face of missing traffic data is required.

Methods to deal with missing data have been studied and applied in different areas [71, 72]. The most common method for imputation is mean or mode substitution. The missing value is replaced by the mean (for continuous data) or the mode (for nominal data). Regression methods are also commonly used methods for missing data imputation. The maximum likelihood method is another approach that can be used.

Over the past decade, dealing with missing traffic data has become a focus of ITS researchers. Algorithms and methods to impute missing traffic data have been applied to works in ITS. Regression methods, which replace missing values with predicted scores from a regression equation, are often used to predict the missing values based on the known points. The work presented in Liu, Sharma and Datla [73] proposes an improved local least squares (LLS) approach to impute the incomplete data. The missing traffic data is replaced by a row average of the known values. Then, the vector angle and Euclidean distance are used to select the nearest neighbors. A regression step is used to obtain weights of the nearest neighbors and the imputation results. Liu, Sharma and Datla [74] also applied the non-parametric regress k-nearest neighbor (k-NN) method to impute abnormal traffic volumes during holiday periods.

The tensor decomposition-based imputation algorithm, which inherits the advantages of imputation methods based on a matrix pattern to estimate missing points, was proposed by Tan et al. [70]. Zhang and Liu [71] introduced a technique called least squares support vector machines (LS-SVMs) to predict missing traffic flow data from loop detectors based on spatio-temporal analysis in urban arterial streets. This approach applies the computational intelligence (CI) technique incorporated with a state space approach in missing traffic data imputation.

As stated earlier, insufficient or missing sensory traffic data can reduce the quality of a TIS so that it cannot disseminate complete traffic information to users. In addition, in the traffic forecast area where complete data is needed to calculate the traffic forecast for route guidance, the ability to accurately predict traffic will be reduced.

Even though some existing works have proposed methods to compensate for missing traffic data, most of these are based on the traffic data of observed road segments, which usually comes from the traffic sensors. Surrounding contexts (that is, data other than traffic data from observed road segments) are rarely taken into account. Moreover, the factors that influence traffic congestion in Bangkok are not widely discussed in the ITS research community. To the best of our knowledge, the current TIS in Bangkok do not apply methods that compensate for missing sensory traffic data.

To solve these problems, the theories and concepts that can inform the development of our artefacts are now identified. Theories focusing on data mining and machine learning, context and context awareness, and ITS, are considered to be relevant to the design of our artefacts. In the next section, we provide background and discussion regarding these theories.

## *2.5 Pervasive Computing and Context Awareness*

This section describes pervasive computing and context aware concepts, which are useful for designing our proposed artefact. Some existing works are also discussed.

### 2.5.1 Pervasive Computing and Ubiquitous Computing

In general, ubiquitous is defined as *existing in anywhere at any time*. Ubiquitous computing was first introduced in the 1990s by Weiser [75] and has also been called pervasive computing. The idea of ubiquity was also addressed by Xerox [76]. Eunyoung, Ryu and Paik [77] summarize the concept as follows:

1) *Service is present everywhere and anytime: Instead of carrying a device with you everywhere you go, a device is available wherever you are.*

2) *It is not the device but an environment: The value of ubiquitous computing lies in the fact that it is a comprehensive environment rather than a collection of services supplied by individual devices.*

3) *The user is not conscious of the device being used: Using a service does not require conscious awareness of the device, thus allowing the user to concentrate on the task at hand.*

4) *Service is TPO (time, place, and occasion) based: The service available matches the situation and need of the user.*

The concept of ubiquitous computing directly affiliates with modern ITS. Modern, transparent devices and services operate in the background so that users do not feel their existence. Traffic sensors, vehicle sensors, display devices and management systems perform together in order to satisfy users and provide transport efficiency, comfort, safety and environmental conservation. The communication between these components is performed with little human intervention.

### 2.5.2 Context Awareness and Situation Awareness

Context can be regarded as "…any information that can be used to characterize the situation of an entity [36]." It ranges from low-level parameters such as time and temperature to highly abstract concepts such as intention and social relationship [37]. The context is what surrounds, and in mobile and ubiquitous computing the term is primarily used in reference to the physical world that surrounds the use of a mobile device [78].

The research by Padovitz [79] defines context as the set of facts, assumptions and predictions along with methods or algorithms of interpreting, discovering, or processing that information. The situations can be inferred by analyzing the contextual information. This hierarchy is illustrated in Figure 2-4.

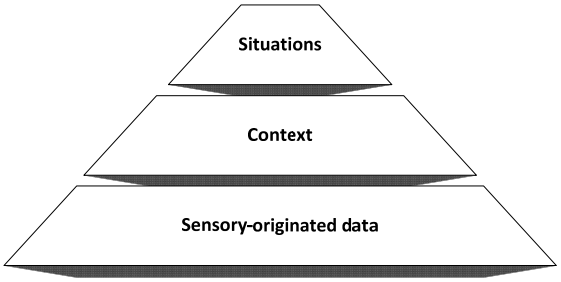



**Figure 2-4: The context-situation pyramid: a three level hierarchy of concepts for modelled information [79]**

A context attribute is defined as any type of data that is used in the process of reasoning about context. A context attribute is associated with a sensor, virtual or physical. The value of the sensor reading at time *t* is the context-attribute value at time *t*. The result from previous processes can become context attributes for later process [79].

Context awareness refers to the idea that computers can both sense and react based on their environment. The computer can react in accordance to certain predefined rules or on the basis of intelligent stimulus. This denotes the capability of the service to utilize context information in order to dynamically adapt its behavior [80, 81].

Situation awareness with respect to our research can be defined as the perception of the system where environment and specific time, space and human behaviours are incorporated [36, 82, 83]. According to the context, an ideal ITS should be able to summarize relationships between attributes in different conditions and then adopt an approach depending upon the specific circumstances.

The concept of context-aware approaches has been widely applied in different areas. For example, a study by Nava-Muñoz and Morán [84] proposed the application of a context-aware system in an aged-care nursing home. Should a resident display atypical behaviour, the system can collate the symptoms based on the physical and mental status of the resident (eg aggression, disorientation, wandering and falls) and notify caregivers. The system can then determine the response based on its decision tree including proximity, criticality of the activity (priority) and attention level.

The context-aware paradigm is relevant to smartphone software developers as well. Floch et al. [85] propose the development of self-adaptive context-aware mobile applications. As a result the MUSIC framework [86] acts as middleware that can gather context information (for example, users' current activities (such as sitting, walking or driving), network properties (such as Wi-Fi signal strength) and GPS location) and then provide this information to applications at run-time so that the applications do not have to create their own context information.

The concept of context awareness is also applied to network and wireless communication. The increasing usage of smartphones and tablets has led to serious load problems for radio networks [87]. A project team described in Proebster et al. [88] applied a context-aware approach to its work. As a result, the team proposed the idea of using information gathered from both handheld devices and Radio Resource Management in order to provide better utilization of the overall network.

While most works that attempt to complete missing data rely only on the traffic data of an observed road segment, we suggest that the concepts of context and context-awareness can be applied to solve the problem of missing sensory traffic data. We can utilize current available contexts in the environment to infer the traffic congestion degree and thus compensate for the missing sensory traffic data. The context to be applied should be context that is related to the observed road segment. For example, the context that makes traffic congestion estimation possible includes incident reports, sensory data, weather data, school break information, day, time, special holiday information, user profiles, cellular phone location, street layouts, building construction and vehicle type. However, we should select only the contexts that have the most effect on the traffic situation. The designed artefact, which is a framework to estimate traffic congestion and is resilient to missing sensory traffic data, should be able to adjust itself to current circumstances. The concept of context awareness, which refers to the idea that computers can both sense and react based on their environment and have the capability to utilize context information, should be incorporated into our design.

In addition to context awareness, we need to have a core process and method to calculate the inferred traffic data from surrounding contexts in order to compensate for

the missing sensory traffic data. We propose that the data mining paradigm is suitable for this purpose.

## 2.6 Data Mining

According to [89], "data mining is the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, by using pattern recognition technologies as well as statistical and mathematical techniques". It often focuses on the automated search of knowledge, patterns or regularities from data that underlie business problems. Different data mining methods can be used to solve problems; for example, estimating probability from historical data, or finding correlations between data sets or variables. The different methods or algorithms can perform the following tasks.

1) Classification or probability estimation. This estimation method attempts to predict the group or class that a record (or an individual) belongs to by creating models based on scoring or probability estimation techniques.

2) Regression or value estimation. The expected result from a regression algorithm is a numerical value of a particular variable for a record or an individual. A model is generated to estimate the value.

3) Similarity matching. The purpose of similarity matching is to identify similar individuals based on data known about them. It usually underlies other data mining tasks such as classification, regression and clustering.

4) Clustering. Based on similarity, individuals are grouped together without predefined purpose. Clustering is useful in the preliminary exploration of existing groups in a data set.

5) Co-occurrence grouping. Co-occurrence grouping is also known as frequent itemset mining, association rule discovery or market based analysis. It uses similarity of objects based on their transactions to assign groups.

6) Profiling. Profiling is based on behaviour description. Profiling is used to establish behavioural norms by characterizing the typical behaviour of an individual, group or population.

7) Link prediction. The algorithm is expected to predict links or connections between data items. For example, Facebook suggests 'friends' based on the strength of the connection.

8) Data reduction. Because a smaller set of data is easier to manage, reducing a large data set to a smaller set - along with the condition that it must still represent the characteristics of the larger set - can make the data mining process more efficient

9) Causal modelling. This algorithm attempts to determine the events or actions that influence other individuals or records by using random controlled experiments or other sophisticated methods to draw causal conclusion from data.

Different tasks are suitable for different situations. However, a business problem may require more than one of the above tasks to find a solution. In order to determine the best formulation, it is important to understand the different types of data mining questions. Usually the process begins with a business problem. The method can be formulated based on these questions whether they have specific purpose or target or not.

The actual data mining implementation involves two main steps: (1) mining the data to find patterns and build models, and (2) using the results of the data mining. Finding patterns and building models in (1) should reflect the business problem or the use of the results in (2). As shown in Figure 2-5, the model is built based on the historical data and data mining process (the upper half of the figure). Then the result, which is the model, will be used with the new data to predict the class value of the model with the probability that the class variable will take on that value (the lower half of the figure).

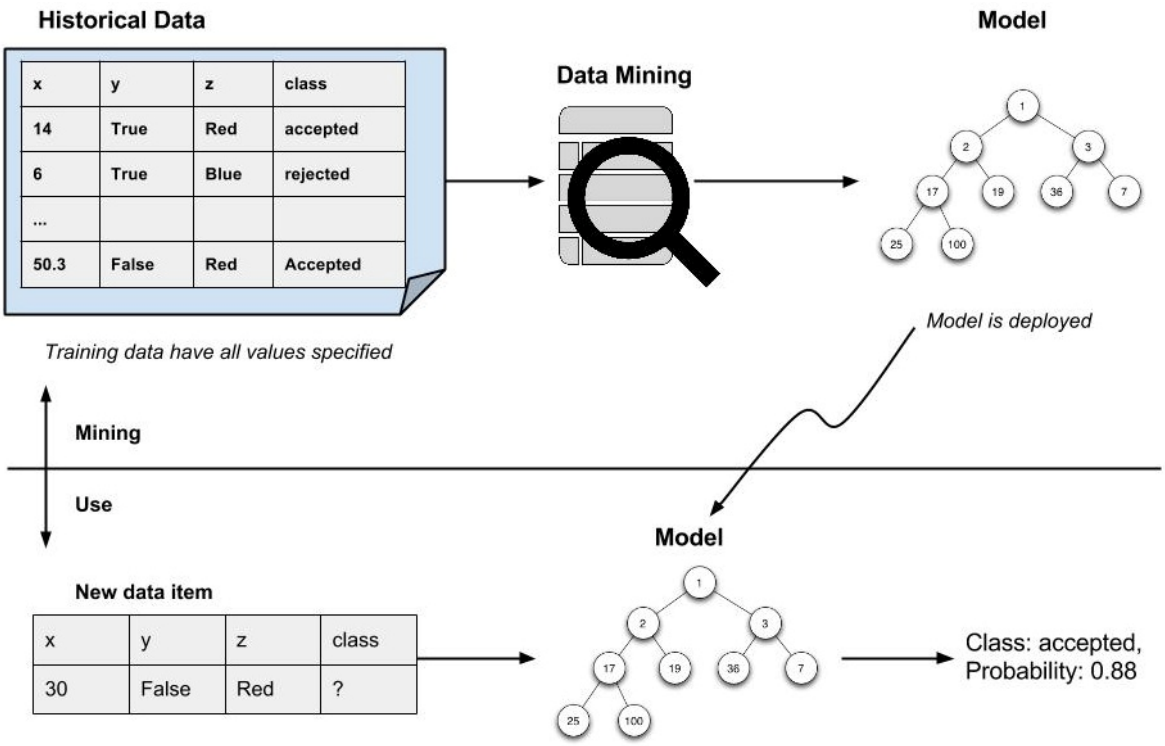



**Figure 2-5: Data mining versus the use of data mining results [90]**

One of the data mining methodology and process models that can be used as a blueprint for conducting a data mining project is CRISP-DM proposed by Shearer [91]. CRISP-DM is a widely accepted methodology and process model. It was designed to provide a generic process model for data mining and was formed by input from a wide range of practitioners and other professionals working in the field. The model breaks down the life cycle of data mining into six phases. Each phase of CRISP-DM is summarized in Figure 2-6 and the tasks and outputs of the CRISP-DM reference model are summarized in Figure 2-7.

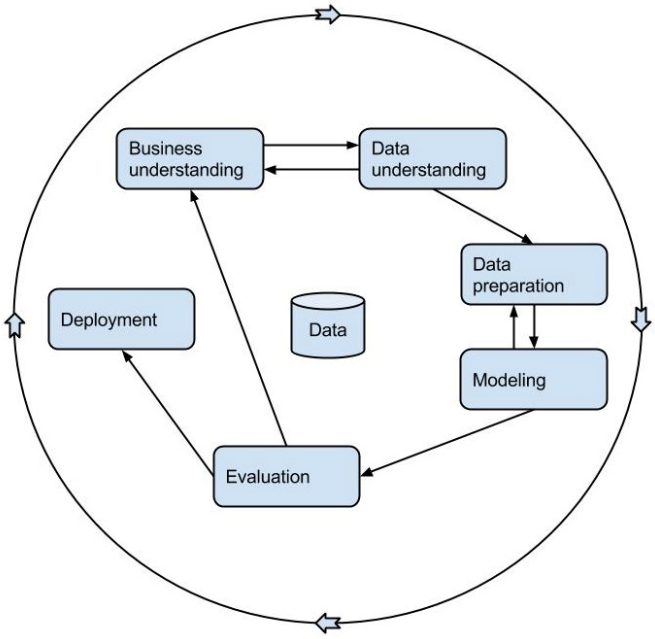



**Figure 2-6: Phases of the CRISP-DM reference model [91]**

| Business Understanding | Data Understanding | Data Preparation | Modeling | Evaluation | Deployment |
|---|---|---|---|---|---|
| **Determine Business Objective**<br>• Background<br>• Business objectives<br>• Business Success Criteria<br>**Assess Situation**<br>• Inventory of Resources<br>• Requirements, Assumptions and Constraints<br>• Risks and Contingencies<br>• Terminology<br>• Costs and Benefits<br>**Determine Data Mining Goals**<br>• Data Mining Goals<br>• Data Mining Success Criteria<br>**Produce Project Plan**<br>• Initial Assessment of Tools and Techniques | **Collect Initial Data**<br>• Initial Data Collection Report<br>**Describe Data**<br>• Data Description Report<br>**Explore Data**<br>• Data Exploration Report<br>**Verify Data Quality**<br>• Data Quality Report | **Data Set**<br>• Data Set Description<br>**Select Data**<br>• Rationale for Inclusion /Exclusion<br>**Clean Data**<br>• Data Cleaning Report<br>**Construct Data**<br>• Derived Attributes<br>• Generated Records<br>**Integrate Data**<br>• Merged Data<br>**Format Data**<br>• Reformatted Data | **Select Modeling Technique**<br>• Modeling Technique<br>• Modeling Assumptions<br>**Generate Test Design**<br>• Test Design<br>**Build Model**<br>• Parameter Settings<br>• Models<br>• Model Description<br>**Assess Model**<br>• Model Assessment<br>• Revised Parameter Settings | **Evaluate Results**<br>• Assessment of Data Mining Result w.r.t. Business Success Criteria<br>• Approved Models<br>**Review Process**<br>• Review of Process<br>**Determine Next Steps**<br>• List of Possible Actions<br>• Decision | **Plan Deployment**<br>• Deployment Plan<br>**Plan Monitoring and Maintenance**<br>• Monitoring and Maintenance Plan<br>**Produce Final Report**<br>• Final Report<br>• Final Presentation<br>**Review Project**<br>• Experience Documentation |

**Figure 2-7: Tasks and outputs of the CRISP-DM reference model [91]**

Data mining has become an essential component of problem solving and been integrated into various fields such as statistics, database, machine learning, pattern recognition and artificial intelligence [92]. The task of data mining can be seen as the

broader process of knowledge discovery of which includes machine learning as one component [73].

In this thesis, we apply the data mining paradigm by using machine learning techniques to learn and build models from massive data for the "modeling" task in Figure 2-6 and Figure 2-7. The details of machine learning and learning algorithms are covered in the next section.

### 2.6.1 Machine Learning Scheme

Machine learning involves programming computers to optimize a performance criterion using example data or past experience. The models are defined based on learning algorithms using training data or past experience. The model may be predictive to make predictions, or descriptive to gain knowledge from data, or both. Machine learning uses statistical theory to build mathematical models, because the core task is making an inference from a sample. In the training process, efficient algorithms are required in order to solve the optimization problem, as well as to store and process the massive amount of data. After a model is learned, the inference is made [93]. If we only depend on existing values we may not obtain any answer apart from that already existing within the data. We thus have to find the relationship between existing data in order to define the rules to forecast the future. The purpose of machine learning is to achieve accurate predictions by gathering sufficient data and then applying learning algorithms. Such a technique to use previous information to forecast the future is not new for humans. Fishermen seek the most effective lure patterns to catch fish and blacksmiths seek patterns in iron purifying. Since then, the data mining approach has been proposed and described in [94] as *"the process of discovering patterns in data. The process must be automatic or (more usually) semiautomatic. The patterns discovered must be meaningful in that they lead to some advantage, usually an economic advantage. The data is invariably present in substantial quantities"*.

Several algorithms that have been designed for machine learning: decision tree learning, association rule learning, neural networks, induction logic programs, support vector machines, clustering, Bayesian networks and support-vector machines. Each approach has been designed and improved to satisfy specific problem domains. These

algorithms can be categorised into two common learning methods: supervised and unsupervised [95].

### 2.6.2 Supervised Learning

The supervised learning methods require training data that includes both the input and the desired results. The term "supervised" denotes that the output values for training samples are already known [96], similar to teachers who supervise learners by providing target information along with a set of examples [90]. Input and output (that is, the correct targets) are fed into the model for the learning process. A supervised learning algorithm then analyzes the training data and produces an inferred function. This inferred function can then be used for mapping new examples. The construction of an appropriate training, validation and test set is crucial. The optimal scenario will allow the algorithm to correctly determine the class labels for unseen instances. Classification and regression are common tasks supported by this type of learning. Some examples of well-known supervised learning algorithms are backpropagation, neural networks, logistic regression, case-based reasoning, naïve bayes classifier, random forests, decision trees learning, bagged trees and boosted trees [97].

**Backpropagation**

Backpropagation is an abbreviation of "backward propagation of errors", a common method of training artificial neural networks used in conjunction with an optimization method such as the gradient descent, Multilayer Perceptron (MLP) [98]. The Multilayer Perceptron or MLP utilizes a supervised learning technique called backpropagation for training the network. The MLP is a modification of the standard linear perceptron. MLP maps sets of input data onto a set of appropriate outputs. MLP consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one [99].

**Case-based reasoning (CBR)**

CBR can be used for classification and regression. It solves new problems based on similar past problems by reusing information and knowledge of that situation. Case-based problem solving is often described as the way humans solve problems. According to Aamod and Plaza [100], this claim is supported by results from

cognitive psychological research. When a problem is successfully solved, the experience is retained in order to solve similar problems in the future. However, when it fails to solve the problem, the reason for failure will be identified. The working cycle of CBR can be described as 1) Case retrieval (finding the best match case), 2) Case adaptation (adapting the retrieved approach to fit the new problem), 3) Approach evaluation (evaluating the approach before or after the adaptation), and lastly 4) Case-base updating (adding the new case to the case base if the approach was verified as correct).

**Decision tree learning**

Decision tree learning is a widely used and practical method. The predictive model is represented by a decision tree which is basically a set of if-then rules. Each node in the tree specifies a test of some attribute of the instance and each branch descending from a node corresponds to one of the possible values for this attribute. An instance is classified by starting at the root node of the tree, then moving down the tree branch corresponding to the value of the attribute. This process then repeats itself for the lower subtree [101].

Decision-tree algorithms include Iterative Dichotomiser 3 (ID3) and C4.5 (the successor to ID3). ID3 is an algorithm invented by Quinlan [102] that generates a decision tree from a dataset. ID3 is the precursor to the C4.5 algorithm, and is typically used in the machine learning and natural language processing domains. The C4.5 was also developed by Quinlan and is an extension of Quinlan's earlier ID3 algorithm. The decision trees generated by C4.5 can be used for classification [103].

**The C4.5 decision tree**

C4.5 [103] is a technique to construct a decision tree (sometimes called a classification model) from a set of training data. The C4.5 algorithm is a statistical-based data mining approach that can be used to build models. It is simple and quick, yet achieves accurate results. It can handle both continuous and discrete attributes and data with missing attribute values. C4.5 builds decision trees from a set of training data using the concept of information entropy. C4.5 also prunes a tree after the tree is completely grown by attempting to remove branches that are not significantly helping the classification.

**The Ripper algorithm**

The use of decision trees in machine learning was discussed and improved by Quinlan [104]. This resulted in pruning strategies that could perform better on unseen and noisy data. As a consequence, the Reduced Error Pruning (REP) technique was proposed to reduce error rates. However, the REP technique was quickly replaced by another improved technique, Incremental Reduced Error Pruning (IREP).

JRip is an optimized version of IREP [105]. JRip comprises a building stage and an optimization stage. The building stage repeats rule growing and pruning until the error rate is less than 50%. The optimization stage optimizes the rule sets obtained from the rule growing and pruning by examining all rules. If residual positives still exist, the building stage again generates rules based on the residual positives [94, 106].

The RIPPER algorithm (Repeated Incremental Pruning to Produce Error Reduction) [105] is an optimized version of IREP. RIPPER purifies rule sets by repeating them as many times as desired. This is reflected in the term *k* of RIPPER, where *k* is the number of repeats. This algorithm is a class-based ordering inductive rule learner. JRip is also a rule-based RIPPER algorithm designed for fast and effective rule induction. As a learner, JRip builds sets of rules identifying classes while minimizing the amount of error [94].

**The Bayesian network**

The Bayesian network is a directed acyclic graph that contains nodes, attributes and directed edges. One example of the popular Bayesian network approach is the Naïve Bayes learning algorithm. Naïve Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes theorem with strong independence assumptions between attributes [107, 108].

### 2.6.3 Unsupervised Learning

Under the unsupervised learning approach, only samples with input values are given to a learning system. In contrast to supervised learning, the model is not provided with the correct result during training. This is similar to the situation where a teacher is not present but the learners form and evaluate a model on their own. The goal of unsupervised learning is to discover "natural" structure in the input data. In machine

learning, the key problem of unsupervised learning is finding the hidden structure in unlabeled data. Because the examples given to the learner are unlabeled, no error or reward signal exists to evaluate a potential approach. The approach can be used in clustering partitioning and labeling sets of data [90, 93]. Examples of unsupervised learning are k-means clustering, hierarchical clustering and hidden Markov models (HMM). The k-means clustering is a method that partitions data into *k* clusters, based on the data's features. Each cluster is represented by its centroid or the center of points in the clusters. Each point is assigned to the cluster whose center is nearest. The goal of this learning is to minimize intra-cluster variance or the sum of squares of distances between data and the corresponding cluster centroid. The advantage of this approach is its speed and simplicity. However, there is no guarantee that the optimal configuration will be found or that the same result will be yielded with each iteration [109].

Hierarchy clustering is a method that builds a hierarchy of clusters with a tree structure based on the distance or similarity between clusters. The tree is called a dendrogram and the objects are joined together from the closest (that is, the most similar) to the furthest (that is, the most different). The method starts building the dendrogram by finding the pair of samples that are the most similar; it then creates the node that joins the pair. The process is then repeated in order to build a complete tree. The key is to decide on how to quantify dissimilarity between two clusters. The techniques to decide this include maximum (or complete) linkage, minimum (or single) linkage or average linkage [110].

A hidden Markov model (HMM) is a statistical model of process consisting of two random variables (O and Y) that change their state sequentially. The model represents probability distribution over a sequence of data. This makes it practical for modeling time series data, speech signals and genomic sequences [111, 112].

## *2.7 Implications for Artefact Development*

The literature review described in this chapter, and interviews with ITS researchers, identified a number of limitations associated with existing TIS and TIS research approaches. These limitations both led to our research questions and guided the direction of this research. Theories and concepts that could inform the development of

our artefacts were reviewed. Theories focusing on data mining and machine learning, context awareness and ITS were all considered relevant. As a result of the review, we determined that our tentative artefact needs to be a framework that can provide traffic data at all times and is resilient to missing sensory data. To overcome missing sensory data, the framework needs to utilize the available surrounding context instead of solely relying on sensory traffic data or the traffic data of an observed road segment. The design of our artefact should enable the artefact to adapt itself to current circumstances. This can be achieved by the artefact automatically adjusting its mode to provide the inferred traffic congestion degree when sensory traffic data is missing. To reach this goal, the data mining process can be used as fundamental to our artefact as we need to build models that can be used to generate inferred traffic data whenever actual sensory data is missing. We need to choose an appropriate machine learning algorithm that is suitable for the type of data and the purpose and requirements of the artefact, including accuracy and processing time. Our artefact should thus apply the context-aware and data mining paradigms to solve the research problems we found.

## *2.8 Chapter Summary*

This chapter reviews existing research and provides the knowledge base relevant to the design of our artefact. The related work and background information in this chapter have demonstrated that most traffic estimation approaches rely heavily on data from sensors of observed road segments. Other types of data sources in the environment are rarely taken into account. Without sensory data, the effectiveness and applicability of these techniques are limited. Yet sensory data can become unavailable due to various reasons. When stationary sensors are used, unsuccessful data transmission and poor weather conditions may lead to data being lost. On the other hand, mobile sensors, such as GPS equipment and mobile phones, can move away from an observed area. Also, due to investment costs, it is not possible to install sensors on every road, especially minor roads where the cost cannot be justified. Absent sensory traffic data can make traffic reports incomplete and may result in subsequent miscalculation of traffic forecasting for route guidance.

Even though existing works propose methods to compensate for missing traffic data, most of these focus on the traffic data of the observed road segment and the

surrounding context is rarely taken into account. Moreover, the studies investigating the factors that influence traffic conditions in Bangkok are scarce and to the best of our knowledge, the TIS in Bangkok do not apply methods to compensate for missing sensory traffic data. Furthermore, minor roads lack sensor infrastructure even though information about these roads would assist road users in route planning. Research to date does not appear to have comprehensively addressed this problem.

To ensure the quality of traffic information, it is therefore necessary to have a TIS that tolerates missing sensory traffic data by adapting itself to the current situation and providing an alternative source of traffic data.

The theories and concepts that can inform the development of our artefacts were collected and discussed in this chapter. Existing theories focusing on data mining, machine learning, context awareness and ITS were all considered relevant. We thus decided to propose a framework that utilizes the available surrounding context to infer the traffic congestion degree in order to overcome missing sensory data. The framework is based on the data mining paradigm so that it can adjust itself to current circumstances. Our research methodology to address the above issues is described in the next chapter.

# Chapter 3 Research Methodology

This chapter presents a summary of the research methodology utilized for this study. It includes a discussion of the research method and design adopted in this research. The primary goal of this research is to design an artefact - The context-aware traffic congestion estimation framework to overcome missing sensory data (CATE framework) - that will help improve the existing traffic information systems in Bangkok, Thailand. To achieve this, we identify context attributes that influence the traffic conditions in Bangkok and also proxy data that approximates and compensates for absent sensory data.

The chapter begins with a discussion of the research questions. These arose as a result of the limitations associated with the existing traffic information systems. The chapter then explains and justifies the adopted research methodology: design science. The research process is described in Section 3.3 and the chapter summary is presented in Section 3.4.

## *3.1 Research Questions*

Traffic congestion is a serious problem for commuters in metropolitan areas. In countries with constrained road infrastructure such as Thailand, Indonesia and the Philippines, the problem is exacerbated.

In order to solve traffic problems, intelligent transportation systems (ITS) (systems that integrate existing and emerging technologies with both information technology and telecommunications to facilitate traffic management and solve traffic problems) have been adopted in many countries, although with different purposes. Knowing traffic conditions in advance helps drivers plan their trips more efficiently. Traffic information systems (TIS) can play a significant role in improving traffic problems. The goal of a TIS is to provide travellers with useful traffic information to assist their route choices. Normally, the TIS involves network infrastructure and sensors to collect necessary traffic data. The data is then processed to produce useful travel

information and is disseminated to users through devices such as electronic boards, websites and mobile phones.

One category of information that assists drivers in making route decisions is the traffic congestion degree. This information is usually calculated from data collected and provided by road side sensors or mobile probes.

As mentioned in Chapter 2, several techniques for traffic state estimation and prediction have been proposed. To date, these approaches have relied mainly on sensors (for example, cameras and loop detectors) deployed in each road segment. However, under certain circumstances, the sensory data may be lost. Three such scenarios are presented in this thesis, as follows.

First, when using static sensors, traffic data may be lost due to a number of reasons. It can cause by unsuccessful data transmission, poor weather conditions or faulty sensors.

Second, if the system relies on mobile sensors such as GPS equipment or cellular phones as traffic probes, these mobile sensors may only be intermittently available. This is because the primary characteristic of mobile sensors results in the abruption of traffic data as the mobile sensor moves. This problem is more severe when mobile phones are used as traffic probes. The number of mobile phones available in each road segment (or even their presence) cannot be guaranteed because mobile phone users move and their movements cannot be anticipated. The loss of sensory data then results in incomplete, and less usable, traffic reports.

Third, in many countries (and especially in developing countries), investment in sensors and infrastructure is limited by a capital budget. Governments normally choose not to invest in traffic sensors for minor roads as the volume of traffic does not justify the investment. We define these kind of roads as “resource constrained”.

In summary, sensory data can become unavailable at particular times due to unsuccessful data transmission, poor weather conditions or faulty sensors. In addition, sensory data may only be intermittently available if the system relies on mobile sensors, or not available at all on resource-constrained roads.

To overcome this problem, the aim of this research is to develop a framework that can deduce traffic relevant data from the surrounding environment that can compensate for the missing sensory data. Traffic relevant data is approximated by using acquirable and readily available context attributes from the environment instead of needing to rely only on sensory data from traffic sensors. This ensures the completeness of the data provided to the real time traffic report service of the traffic information system at all times. To guide the research process, two research questions were posed at the beginning of the research.

***Research Question 1:***

***What information can be collected from sources other than the road sensors to approximate missing sensory data?***

In order to address the first question it is necessary to understand the factors that affect traffic conditions. We began by literature review and interviewing experts who are researchers in intelligent transportation system in order to generate an initial list of influential factors. We then conducted a web based survey with road users to identify their perception of influential factors. The survey results confirmed our selected list of influential context attributes. We then refined the list of influential context attributes to obtain the reduced set of influential context attributes. The refinement is based on the survey result on the perception of Bangkok traffic and the knowledge gain from running an experiment. This reduced set of influential context attributes was then used to shape our initial framework into the final artefact.

***Research Question 2:***

***How would data collected from other sources be processed to approximate the missing sensory data?***

To address the second research question, it is necessary to understand intelligent transportation systems (ITS), traffic information systems (TIS), traffic dissemination systems, traffic congestion estimation and prediction methods, data mining, machine learning algorithms, and the context awareness concept by reviewing associated theories, methodologies and existing approaches. Such knowledge is described in Chapter 2. This understanding underpins the artefact design. Once the tentative framework was designed, we implemented the program and simulation according to the designed framework to evaluate, refine and conclude the proposed solution to the second research question.

This chapter explains how these research questions were addressed and how the research activities were justified and organised in the design science research paradigm.

## *3.2 Design Science as the Research Approach*

Design normally involves the creation of something new that does not exist in nature. Design science can be regarded as an alternative to the natural science approach. Natural science is normally concerned with explaining how and why things are [113]. Design science, in contrast, is concerned with devising artefacts to attain goals [114]. March and Smith clearly outline the difference between natural and design science in [115]. The purpose of natural science is to understand reality and develop sets of concepts for characterizing the phenomena. Its products are evaluated with norms of truth or explanatory power. The end result of this process is new theories that provide more detail, are more encompassing and have more accurate explanation prowess. In other words, where natural science tries to improve our understanding of reality, design science seeks to create things that serve a given human purpose. Hevner et al [33] later explained that design science can be described as a problem solving activity that focuses on new technologies. Design science addresses research through the building and evaluation of an artifact as design to meet the identified research need. The goal of design science research is utility.

Our research uses the design science research cycle proposed by Hevner [32, 33] as the overarching framework. Under this umbrella, we follow the design science

research steps proposed by Vaishnavi and Kuechler [34, 35]. Our research steps and process is described in section 3.3. In addition, the research is guided and assessed according to the guidelines for design science introduced by Hevner et al. [33]. (summarized in Table 3-1). The adaptation of Hevner's seven guidelines [33] to assess the quality of our research is concluded in Chapter 8.

**Table 3-1: Guidelines for judging design science research quality by Hevner [33]**

| Guideline | Description |
|---|---|
| Guideline 1: Design as an Artefact | Design science research must produce a viable artefact in the form of a construct, a model, a method, or an instantiation. |
| Guideline 2: Problem relevance | The objective of design science research is to develop technology-based approachs to important and relevant business problems. |
| Guideline 3: Design evaluation | The utility, quality, and efficacy of a design artefact must be rigorously demonstrated via well-executed evaluation methods. |
| Guideline 4: Research contributions | Effective design science research must provide clear and verifiable contributions in the areas of the design artefact, design foundations, and/or design methodologies. |
| Guideline 5: Research rigor | Design science research relies upon the application of rigorous methods in both the construction and evaluation of the design artefact. |
| Guideline 6: Design as a search process | The search for an effective artefact requires utilizing available means to reach desired ends while satisfying laws in the problem environment. |
| Guideline 7: Communication of research | Design science research must be presented effectively to both technology-oriented and management-oriented audiences |

In addition, Hevner proposed a conceptual framework for design science paradigms in [33] (shown in Figure 3-1). The environment defines the problem space in which reside the phenomena of interest. It is composed of people, organizations and their existing technologies. The knowledge base comprises foundations and methodologies. Prior research and results from others provide foundational theories, frameworks, instruments, constructs, models and methods. The methodologies provide guidelines

used in the evaluation phase. Design science draws from a vast knowledge base of scientific theories and engineering methods that provides the foundations for rigorous design science research. The rigor is achieved by applying existing foundations and methodologies appropriately. Methodologies provide guidelines used in the evaluation phase [33].

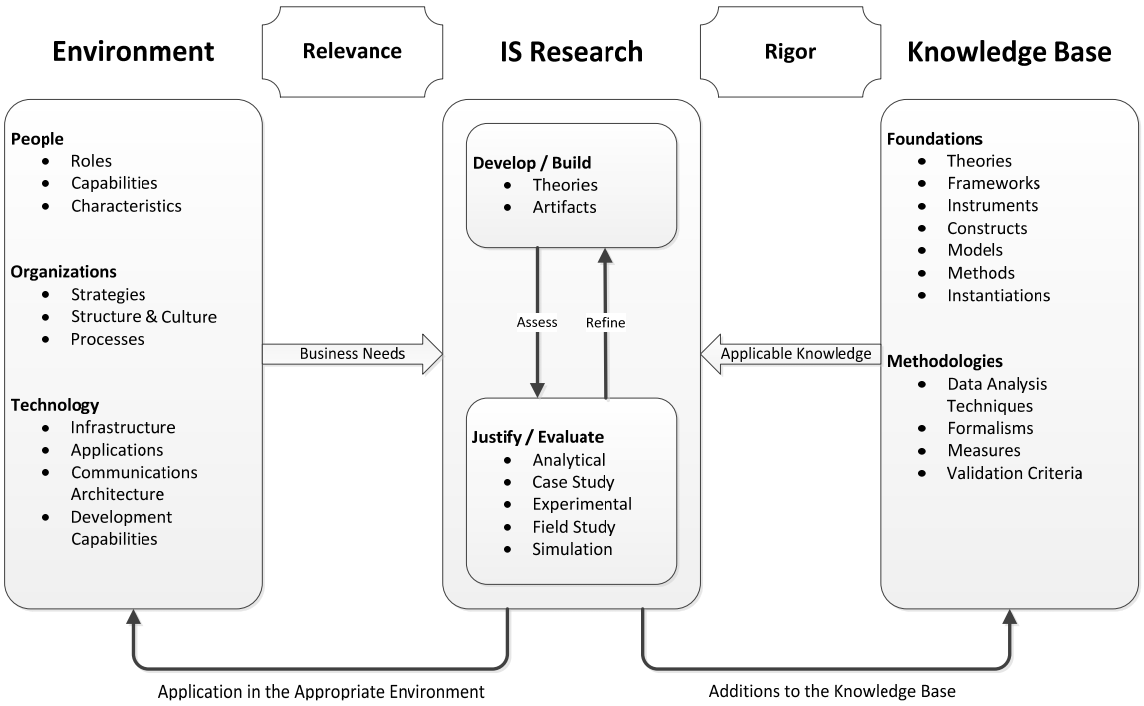



**Figure 3-1: Information system research framework [33]**

The work in [33] does not propose a detailed process for performing design science research. Hevner later proposed a key insight - the conceptual framework for design science in [32] -that clearly differentiates design science from other research paradigms (shown in Figure 3-2). It focuses on three inherent research cycles: the r*elevance cycle*, the *rigor cycle* and the *design cycle*. The *relevance cycle* bridges the contextual environment of the research project with the design science activities. The *relevance cycle* inputs requirements from the environment into the research and introduces the research artefacts into environmental field testing. The *rigor cycle* connects the design science activities with the knowledge base of scientific

foundations, experience and expertise that informs the research project. The *rigor cycle* provides grounding theories and methods along with experience and expertise, from the foundation knowledge base, into the research. It also adds new knowledge obtained from the research to the growing knowledge base. The *design cycle* is the heart of any design science research project. This cycle is the hard work of design science research. This cycle of research activities iterates between the construction of an artefact, its evaluation and subsequent feedback to refine the design further [32].

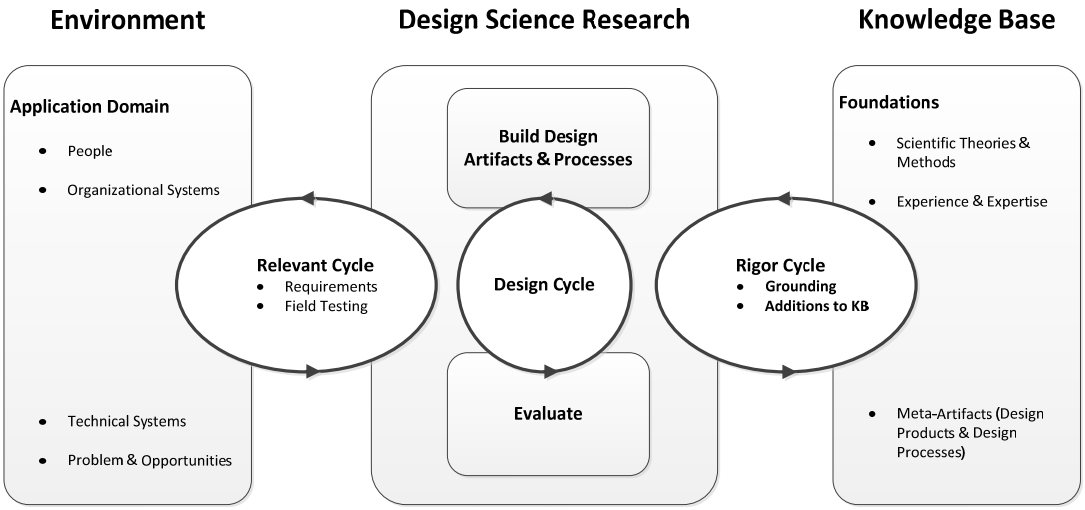



**Figure 3-2: Design science research cycle [32]**

The Heverner's works in [33] and [32] do not suggest the type of output design science might produce. March and Smith's widely cited paper [115] has proposed four general outputs for design science research: (1) constructs, (2) models, (3) methods and (4) instantiations. Table 3-2 summarizes the outputs that can be obtained from a design science research effort [34]. In our project, our artefact comprises the methods that solve the problem of missing sensory traffic data, which currently renders traffic reports in Bangkok incomplete.

**Table 3-2: The outputs of design science research**

| Output | Description |
|---|---|
| Constructs | The conceptual vocabulary of a domain |
| Models | A set of propositions or statements expressing relationships between constructs |
| Methods | A set of steps used to perform a task — how-to knowledge |
| Instantiations | The operationalization of constructs, models, and methods |
| Better theories | Artifact construction as analogous to experimental natural science |

Once the artefact is designed, it must be evaluated with appropriate metrics. According to [33], the evaluation of designed artefacts typically uses methodologies available in the knowledge base. The selection of evaluation methods must be matched appropriately with the designed artefact and the selected evaluation metrics. Typical evaluation methods are summarized in Table 3-3.

**Table 3-3 : The evaluation methods [33]**

| Evaluation Method | |
|---|---|
| Observational | Case Study |
| | Field Study |
| Analytical | Static Analysis |
| | Architecture Analysis |
| | Optimization |
| | Dynamic Analysis |
| Experimental | Controlled Experiment |
| | Simulation |
| Testing | Functional (Black Box) Testing |
| | Structural (White Box) Testing |
| Descriptive | Informed Argument |
| | Scenarios |

Each of the evaluation methods presented in Table 3-3 are suited to particular situations. The case study is best used to study artefacts in depth while a field study can be employed to monitor the use of artefacts in multiple projects. Static analysis can be used to examine the structure of an artefact for static qualities (for example,

complexity). The optimization evaluation method demonstrates the inherent optimal properties of artefacts or provides optimality bounds on artefact behaviour. Dynamic analysis is used to study in-use artefacts for dynamic qualities (for example, performance). The experimental method includes controlled experiments, where artefact are studied in controlled environments for particular qualities (for example, usability) and simulation, where artefacts are executed with artificial data. The testing method involves functional (black box) testing, which executes artefact interfaces to discover failures and to identify defects while structural (white box) testing performs coverage testing of some metric in the artefact implementation. The informed argument uses information from the knowledge base (for example, relevant research) to build a convincing argument for the artefact's utility. Finally, the scenarios method involves the construction of detailed scenarios around the artefact to demonstrate its utility [33]. The evaluation method selected for evaluating our proposed artefact is the experimental method by implementing the program that simulate the real situation using real data and evaluate the proposed framework.

We choose design science research method as the research methodology because it offers specific guidelines for evaluation and iteration within research projects. Hevner's guidelines [33] help researchers to conduct, evaluate and present design science research to ensure the quality of the research and the designed artefact. The iteration within research projects using the design science research methodology [34] offers a more flexible method to address our research problems than traditional research methodology. The results from the early evaluation of our artefact, new problem awareness and additional knowledge can be used to improve the artefact until we obtain the final iteration of the design cycle and the final artefact. Further, in addition to the final artefact, the knowledge and understanding gained from the design cycle can also contribute to the knowledge encapsulated within an ITS. These iteration tasks are explained in Chapter 4, Chapter 5, and Chapter 6.

Another appeal of design science research is that it is distinguished from routine design by the production of interesting (to a community) new knowledge. Design science research requires the creation of an innovative, purposeful artefact for a special problem domain. The artefact must be evaluated to ensure its utility for the specified problem. The artefact must either solve a problem that has not yet been

solved or provide a more effective solution [33, 34]. Our research project focuses creating a new artefact that could solve the problems of existing traffic information systems and provide a more effective solution using an improved knowledge base. Our research is congruent with the aims of design science; therefore we used a design science approach to achieve our goal.

## *3.3 Research Process*

This research has adopted the design science research cycle proposed by Hevner [32, 33] as its overarching framework and the design science research steps proposed by Vaishnavi and Kuechler [34, 35] for the research process. In addition the research is guided and assessed according to the guidelines for design science introduced by Hevner [33] to ensure that the project addresses the key aspects of design science research.

In 1990, Takeda et al. [116] analysed the reasoning that occurs in the course of a general design cycle (GDC). Vaishnavi and Kuechler extended this analysis and applied the cycle specifically to design science research [34, 35] (illustrated in Figure 3-3).

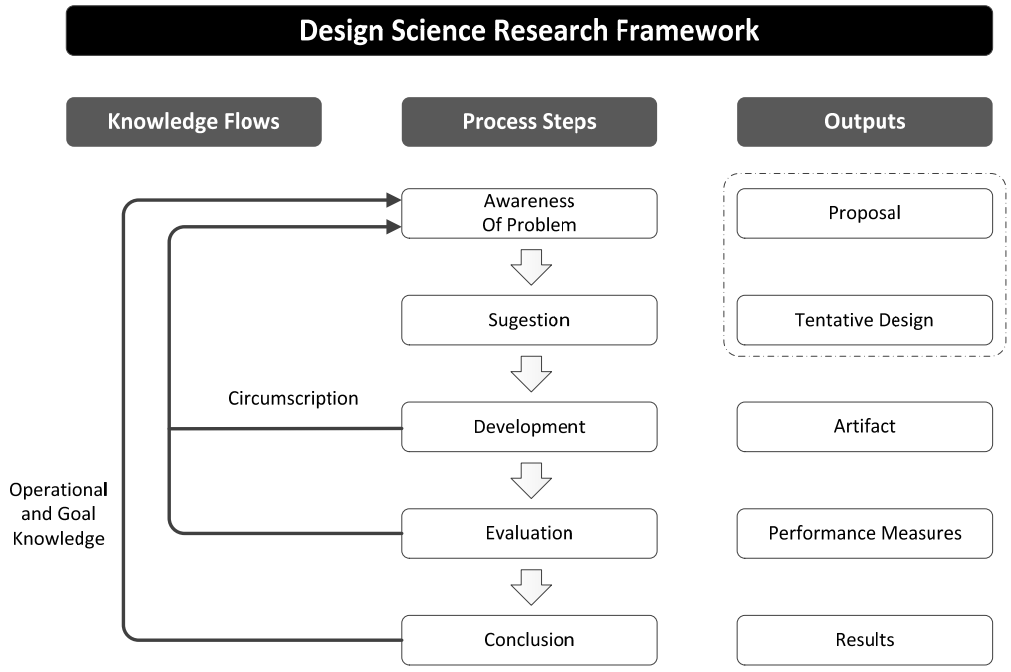



**Figure 3-3: The process steps of design cycle [34]**

According to the general methodology of design science research depicted in Figure 3-3, all design begins with an awareness of a problem and its identification and definition. The next stage is a preliminary suggestion for a problem solution that is drawn from existing knowledge or theory concerning the problem area.

Our research process is illustrated in Figure 3-4 and also corresponded to the general methodology of design science research. In addition to introducing the artefact, we also performed further research to provide guidelines for improving traffic estimation systems and traffic information dissemination systems. We investigated the potential use of social networks for traffic information systems in Bangkok in order to improve the efficiency of traffic information systems. We also investigated Bangkok road users' need for traffic information on resource-constrained roads and their preferred communication channels for this information in order to assist with traffic information dissemination services.

In Figure 3-4, the blue elements in the middle of the diagram are the process steps. Each process step is associated with a task(s), illustrated in the gray elements on the right of the diagram. The output of each process is presented in the orange elements on the left of the diagram.

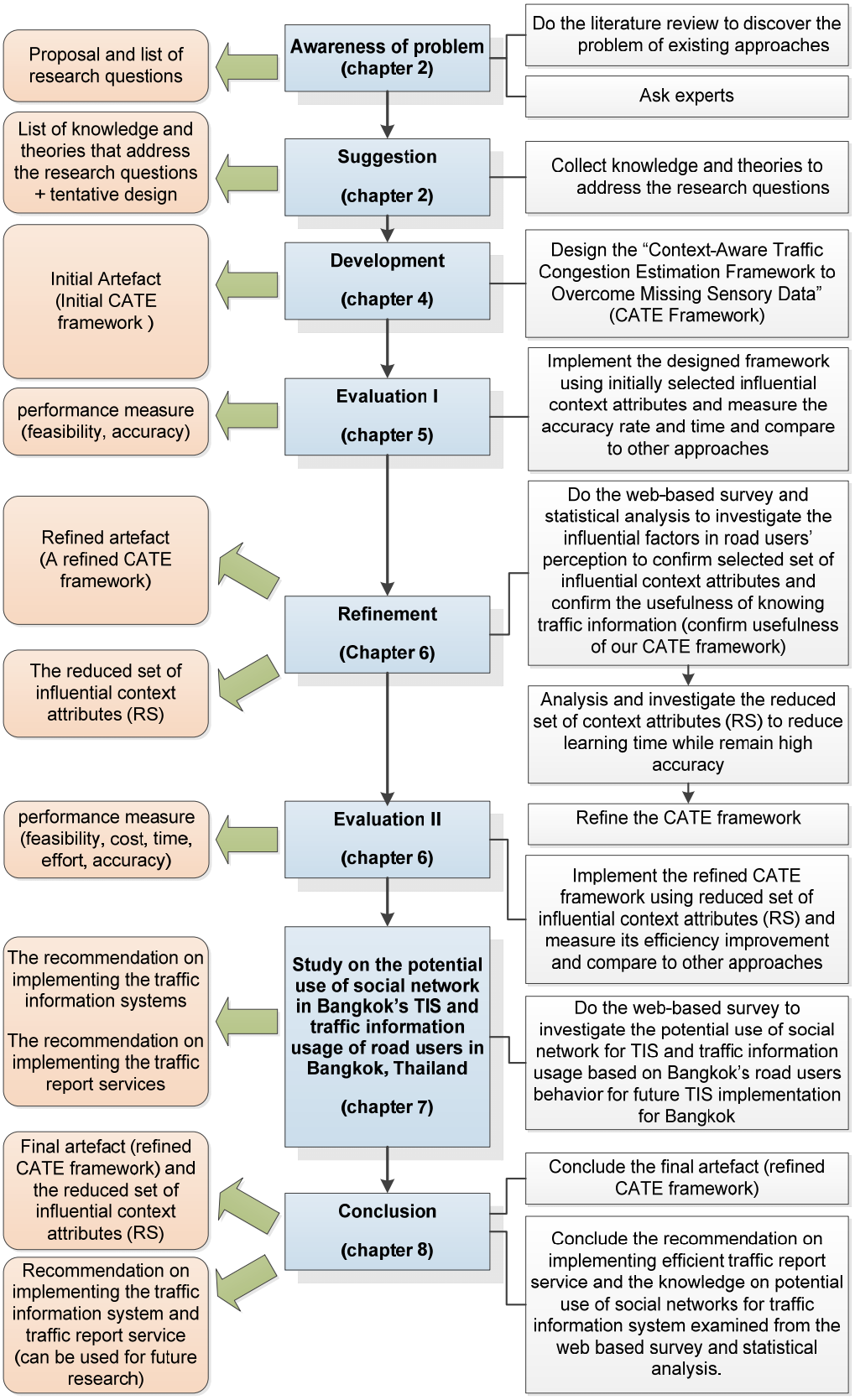



**Figure 3-4: Research Process**

From Figure 3-4, it can be seen that each research phase involved a number of research activities. These activities are introduced in the following sections.

- Conducting the literature review and analysis (see Section 3.3.1). This involved justifying the selected literature and investigating the limitations of existing approaches.

- Collecting the catalogue of related theories (see section 3.3.2). This involved justifying the theories to be used for addressing the research questions.

- Designing the proposed framework (see section 3.3.3). This included justifying the design of the proposed framework under three scenarios (static sensors lacking sensory data, intermittent availability of mobile sensory data and small roads with no sensor infrastructure) and designing the algorithm, data preparation methods and instruments.

- Performing the implementation according to the designed artefact to evaluate the proposed framework (see Section 3.3.4 and section 3.3.6). The first evaluation is an evaluation of the initially proposed framework on the feasibility and accuracy against other approaches (before refinement) while the second is an evaluation based on the accuracy, time, and cost of the refined framework against the initial framework and against other approaches.

- Refining the framework (see section 3.3.5). This involved determining the reduced sets of influential context attributes for efficient traffic information systems. Analysis of these sets was used to refine the proposed framework.

- Undertaking quantitative research and statistical analysis. This involved a web based survey to capture data and statistical analysis to yield the confirmation that guide selecting the reduced set of influential context attributes which is used for our refined framework. The survey result also yields recommendations for improving the traffic information system and for improving the traffic information dissemination service in the future.

### 3.3.1 Awareness of Problem

An awareness of an interesting problem can come from new developments in industry or in a reference discipline. Reading in an allied discipline may also provide an opportunity for an application of new findings. The output of this phase is a proposal for a new research effort. In our study, both a literature review and interviews with experts were conducted to discover the issues associated with existing traffic information systems. The literature review is presented in Chapter 2.

Based on our investigation into existing systems, the literature analysis and the interviews, a number of limitations associated with existing traffic information system were identified. Consideration of these limitations leads to development of the research questions. The questions, and the research proposal, can thus be considered an output of the Awareness of Problem phase.

### 3.3.2 Suggestion

The Suggestion phase follows the proposal and is intimately connected with it. The proposal is an output of the Awareness of Problem phase whereas a tentative design is an output of the Suggestion phase. In the Suggestion phase, theories and concepts that might inform the development of one or more artefacts are referenced. The Suggestion phase involves iterative processes that seek to identify specifications of the design artefacts at a high level.

In order to develop an efficient framework, a number of alternative scientific theories and methods may be useful. Existing theories focusing on data mining and machine learning [94, 103, 117], context awareness [36, 118, 119] and the knowledge base of intelligent transportation systems [47, 77, 120] were considered to be relevant. The list of related knowledge and theories including the suggestion about tentative design of artefact are discussed in Chapter 2.

### 3.3.3 Artefact Development

Vaishnavi and Kuechler stated in [34] categorized the pattern to develop an artefact. Those patterns are Theory Development, Approaches for Building Theory, Hermeneutical and Inductive Approach, Incremental Theory Development, Problem

Space Tools and Techniques, Research Community Tools and Techniques, Empirical Refinement, Easy Approach First, Elegant Design, Divide and Conquer with Balancing, Hierarchical Design, Building Blocks, Sketching Approach, Emerging Tasks, Modeling Existing Approachs, Combining Partial Approachs, Static and Dynamic Parts, Simulation and Exploration, Interdisciplinary Approach Extrapolation, Different Perspectives, General Approach Principle, Abstracting Concepts, Using Surrogates, Using Human Roles, Integrating Techniques, Technological Approach Exemplars, and Means-End Analysis.

Reddy, Finger, Konda, and Subrahmanian stated that as artefacts are designed, knowledge is accumulated gradually. As this knowledge is organized and reused, design and design processes are continually refined [72]. The design of our artefact was generated based on compound theories selected in the suggestion phase. The tentative design was developed in this phase. Elaboration of the tentative design into a complete design requires creative effort. The techniques for implementation vary, depending on the artefact to be constructed. The design of our artefact is presented in Chapter 4 and the final artefact after refinement is presented in Chapter 6.

### 3.3.4 Evaluation I

Once constructed, the artefact is evaluated according to criteria that are always implicit and frequently made explicit in the research proposal generated through the awareness of problem phase. In this study, the evaluation comprised two evaluations. Evaluation I concerned the framework before refinement. At this point, we chose to evaluate using simulation because it complemented the designed artefact and the evaluation metrics. The evaluation was based on the feasibility, efficiency, and accuracy of the designed framework. After further investigation into influential factors in order to confirm our selected list of context attributes combining with knowledge gain from evaluation I, and analysis of a reduced set of context attributes, we refined the proposed framework and evaluated it again in Evaluation II.

### 3.3.5 Refine

The investigation into the factors influencing Bangkok's traffic employed quantitative research through a web based survey. The research involved the statistical analysis of

survey results and the analysis of a reduced set of influential context attributes (based on the results from Evaluation I). The outputs of the investigation were used to improve the design artefact. We refined the proposed framework and evaluated it again in Evaluation II. The refined framework is presented in Chapter 6.

### 3.3.6 Evaluation II

Evaluation II was an evaluation of the artefact after refinement. The refined framework was judged on accuracy against the initial framework and against other approaches. It is also evaluated on the required processing time reduction and cost reduction to demonstrate the improvement. In this step, we validated the accuracy of the artefact using the same methods employed in Evaluation I, with the additional comparison of the artefact before and after refinement including comparison with other approaches. Evaluation II is further described in Chapter 6.

The purpose of Evaluation II was to test the improved inference technique and to evaluate performance improvement. The accuracy and improvement in processing time were evaluated in Chapter 6. The reduction in resources required and cost were also discussed.

In addition to designing the artefact, we performed further research to derive recommendations to improve the traffic information system and the traffic information dissemination system in Bangkok. This research involved statistical analysis of results from a web based survey. The aims of the survey were to determine the traffic information needs of Bangkok road users and the preferred media channels to disseminate this information to users. The survey also collected data regarding the potential use of social networks in Bangkok's traffic information system with the view of determining whether this data could be used to improve both traffic information system and traffic dissemination service in Bangkok. The results from this part of the study are presented in Chapter 7 before the conclusion in Chapter 8.

### 3.3.7 Conclusion

The conclusion phase is the final phase of the research effort. The output of the conclusion phase in our research was the final framework and the recommendations for a reduced set of influential context attributes. The recommendations for

implementing the traffic information system and the traffic report service in Bangkok were additional outputs for this phase. The conclusion phase (Chapter 8) consolidates the results of our research and the knowledge gained. The assessment of the current study using the design science research guidelines introduced by Hevner [33] is included in this chapter. Future research is also discussed.

## *3.4 Chapter Summary*

This chapter has summarized the research methodology and processes adopted in this project. The chapter began by presenting the research questions and establishing the utility of the methodology to address the research questions. In addition, the chapter introduced the concept of design science and explained its suitability for this research. Each process within the design science research method was presented. This included Awareness of Problem, Suggestion, Artefact Development, Evaluation I, Refinement, Evaluation II and Conclusion.

This chapter has explained our research methodology approach. The next chapter explain our proposed framework, a context-aware traffic congestion estimation framework to overcome missing sensory data or the CATE framework.

# Chapter 4 A Context-Aware Traffic Congestion Estimation Framework to Overcome Missing Sensory Data (CATE)

Real time traffic information dissemination is an important component of intelligent transportation systems (ITS) to assist drivers in journey planning. Traffic information can be disseminated through methods such as electronic boards, websites, mobile phones and other devices. Among the various categories of traffic information, the traffic congestion level is simple and easy to understand, yet significantly useful for travellers. The congestion level, or traffic congestion degree, is useful for trip planning. Knowledge of traffic congestion allows drivers to make more informed decisions, yielding benefit for individuals. The same information also contributes to better traffic management and control. In addition, the traffic congestion degree can be used as a context attribute or parameter to facilitate a more accurate travel time calculation rather than static information, such as the journey distance and road speed limits, can offer. The traffic congestion degree can also be used to derive other traffic information.

Chapter 2 described several existing techniques for traffic state estimation and prediction. Most existing traffic state estimation approaches rely on data from sensors (which can be stationary, mobile or a combination of both) dedicated to an observed road segment. Other types of data sources in the environment are rarely taken into account. In addition, an attempt to estimate the traffic conditions in sensorless road segments or when sensory traffic data is missing has not been made in most works. Our approach is different. We focus on utilizing discoverable context rather than focusing on traffic data from sensors because it is possible that the sensory data could be absent due to unsuccessful data transmission. For example, a camera may provide

data that cannot be interpreted under poor weather conditions, or the camera or road side sensors may be broken.

Another approach is to use mobile sensors. However, this approach also has limitations. The mobile sensors are likely to move, leading to the abruption of traffic data. This problem may become more severe when using mobile phone as mobile sensors. The mobile signal may not be reliable in some geographical locations. The number of available mobile phones, or even the presence of a mobile phone, in a road segment may be uncertain because phone users are mobile and determining their availability in advance is infeasible.

Our approach, the CATE framework, is a novel framework that overcomes these restrictions. It compensates for the missing traffic data by using, additionally, inferred traffic information. This guarantees that the traffic data can be provided at all times, even when:

- sensor data is lost due to unsuccessful data transmission or faulty or broken roadside sensors at particular times [121];
- sensor availability is intermittent [122];
- no sensor data is available (such as on minor roads lacking sensor based infrastructure) [123].

While our approach uses sensory traffic data, it also utilizes acquirable context information (for example, date, time, school holiday information, traffic on connected road segments and weather data) and utilizes context reasoning. With this proposed framework, it becomes possible to estimate the traffic conditions of sensorless road segments. The rest of the chapter describes the CATE framework.

## *4.1 Dealing with Resource Constrained Roads*

Our approach is designed to support existing traffic report services that use sensory data as the data source. The real time traffic state report can assist drivers in choosing routes and avoiding blocked roads. In general, the traffic report services obtain the traffic data from assigned sensors, which can be either stationary or mobile. Stationary sensors (such as cameras and inductive loops) are installed along roadways. Mobile

sensors (such as mobile phones and GPS equipment) do not need fixed installation and can move flexibly. To make efficient route choices, drivers require complete traffic information from all roads. In order for drivers to obtain this continuous traffic information, the sensors must send data to the system at determined time periods.

In some periods, sensory data may not be available for some road segments and these road segments become "*resource constrained roads*". This leads to missing traffic data for particular road segments at particular times, thus discontinuing the useful traffic information available to drivers.

In addition, under insufficient road infrastructure environments (for example, in developing countries), the traffic management organization or public sector generally does not invest in traffic sensors for minor roads. However, traffic information for small roads may still be necessary for better route planning by drivers. Summarily, our approach can address three categories of problems, which represent three cases as follows:

- the uncertainty of roadside sensors;
- the intermittent availability of mobile sensors; and
- sensorless minor roads.

Each case is elucidated as follows.

### 4.1.1 Uncertainty of Roadside Sensors

Traffic data in traffic information systems (TIS) can be obtained from various sources. The traffic sensor is the most extensively used source. Different approaches deploy various types of sensors for travel time estimation. A sensor comprises hardware with assisted software that detects vehicles and converts the information into traffic flow data. Sensors that act as traffic data collectors can be both stationary and mobile. We can categorize stationary sensors into intrusive detectors (installed in the roadway) and non-intrusive detectors (installed above or beside the roadway) [10].

In the case of roads with stationary sensors installed, when a sensor is broken or the communication to the data centre is lost, the situation of such a road is defined as

resource constrained because the necessary resource - the sensory traffic data - is missing.

### 4.1.2 Intermittently Available Mobile Sensors

Sensors provide useful traffic data for the TIS to interpret and report to users. Nowadays, mobile phones are not only traffic information receivers, but can also become sensors to provide traffic data to traffic systems. The proposed framework is applicable both to traffic report services that rely on stationary traffic sensors and also to systems that employ mobile sensors as the source of traffic data.

While mobile sensors enable the generation of data that would not otherwise be produced, this same mobility leads to irregular and unreliable dissemination of traffic data on observed road segments due to the abruption of traffic data sources. This problem will be more severe when mobile phones are used as traffic probes. The number of mobile phones available in each road segment fluctuates with the movements of mobile phone users, and it is not feasible to determine in advance the movement plans of each mobile phone user. At times, no mobile phone will be available to act as the mobile sensor for a particular road segment. Figure 4-1 represents the scenario where the mobile sensors are intermittently available for a particular road segment. For example, at time *T1*, all road segments have mobile sensors to provide traffic data. However the mobile sensors move such that at time *T4* some road segments have no mobile sensors, and thus no traffic data can be provided for these road segments.

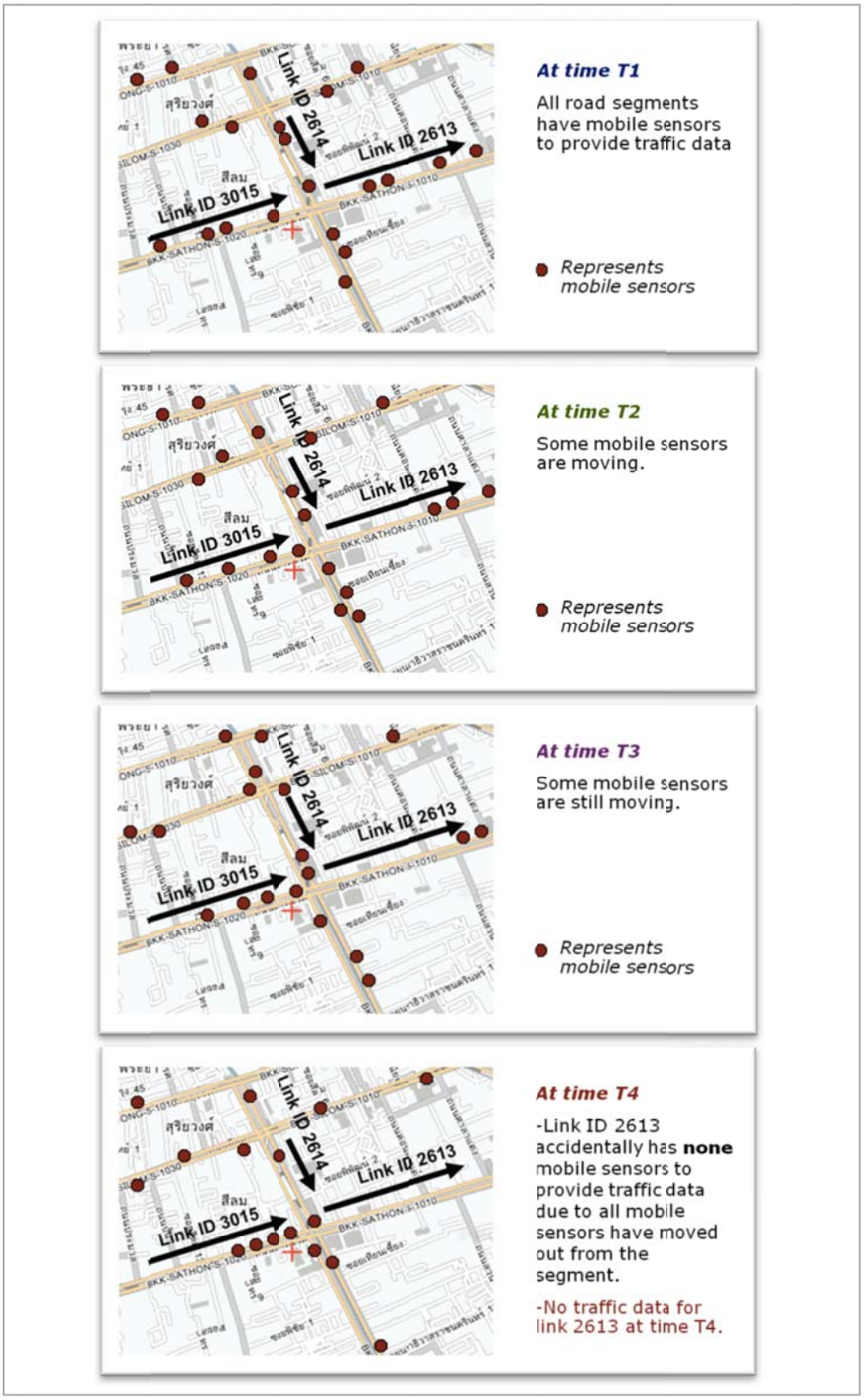



**Figure 4-1: Intermittently available mobile sensors scenario**

Our approach, initially presented in [122], can overcome this restriction. While most traffic congestion or traffic flow estimation approaches rely only on sensory data and

do not consider the surrounding context, our approach is different. We focus on utilizing discoverable context as well as the sensory data of an observed road segment. To compensate for any missing mobile sensory data, our method employs traffic data extracted from discoverable context attributes.

### 4.1.3 Sensorless Small/Minor Roads

A small or minor road refers to a side street branching off a major street. In Bangkok, the minor roads are called “soi”. The sois are usually numbered and are referred to by the name of the major street they intersect, followed by the number: for example, "Soi Sukhumvit 4". The road network in Bangkok consists of many small roads connecting the main roads. In Bangkok, commuters use small roads for their daily travel. However, like many developing countries, Bangkok’s traffic management organization does not normally invest in traffic sensors for minor roads as it is not worthwhile. Yet the traffic information for minor roads would facilitate better route planning. We proposed a approach to address this problem in [123].

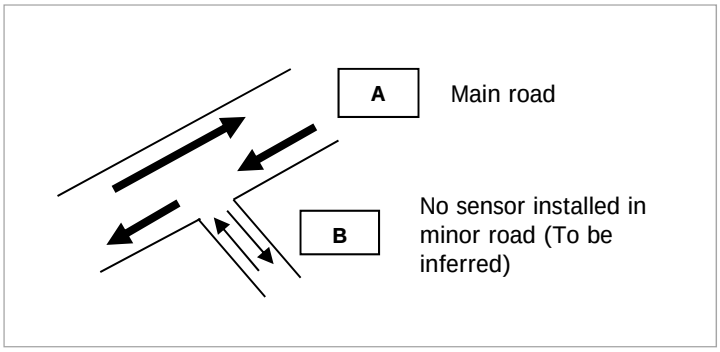



**Figure 4-2: No sensor infrastructure (minor road)**

Figure 4-2 illustrates a small road without sensor infrastructure (B) connected to a main road equipped with sensors (A). The main or major road is a road that commonly has a sufficiently high volume of traffic to make the investment in traffic sensors worthwhile, while the minor road does not have the same volume of traffic. Under this scenario, traffic data for small road B is not available as it has no dedicated sensor. However, we can still acquire other contextual information such as the weather, day, time, traffic congestion degree of the connected road and school vicinity.

To compensate for the missing traffic congestion information, we propose to use the acquirable surrounding contexts to infer a traffic congestion degree for sensorless minor roads instead of relying on the traffic data from dedicated traffic sensors. Our approach is thus cost efficient and suitable for road environments lacking sensor infrastructure. It also aligns with the concept of "anytime", "anywhere" and "anything" for ubiquitous intelligent traffic information system [77]. The proposed approach thus achieves the availability, transparency, seamlessness and awareness goals of ubiquitous IT services.

To solve the problem mentioned in Section 4.1, we propose to use the available surrounding contexts to infer the traffic congestion degree instead of relying on the traffic data from dedicated traffic sensors. To achieve this, we propose to include a context attribute extractor in the framework. We describe context attributes and how a context attribute extractor works in the next section.

## *4.2 Context Attributes and the Context Attribute Extractor*

Before using context to determine the state of traffic, we have to extract the context to a context attribute. The context attribute extractor extracts the context attributes from historical data and available contexts at the required inferring time as illustrated in Figure 4-3. For instance, the rain context attribute is extracted from a weather log while the traffic congestion degree is extracted from a traffic log. Day and time information can be extracted from the traffic and weather logs for historical data and from the current system clock in real time.

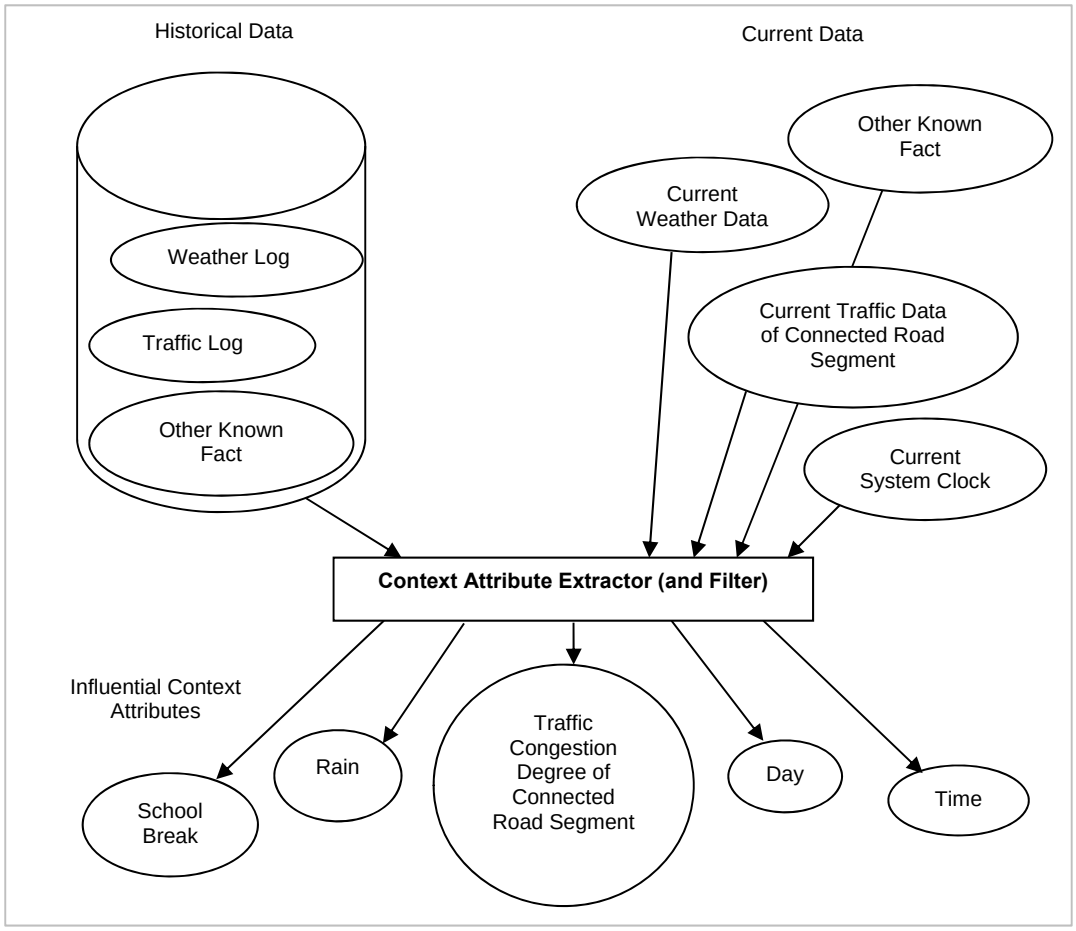



**Figure 4-3: Context attribute extractor**

The context attribute extractor module acts as a data pre-processor to transform acquirable contexts into a format suitable for both learning and estimation processes. The contexts (for example, the weather log, traffic log or known fact) are usually not directly applicable and must be changed into a suitable format before being used. Moreover, the attributes to be extracted and their formats must be carefully analysed to avoid wasting computational and human resources in the learning process. Too much detail may result in excessive learning processing times and estimating processing times without a significant improvement of accuracy. The numeric data should thus be converted into a format suitable for the machine learning algorithm and the output target. For instance, rain data in a weather log is converted from numerical format (millimetres) to "T", "S", "M", "H", or "VH". The log time in the traffic log is extracted into "date" and "time". We also divided "time" into 24 periods and "day"

into 7 different values. Table 4-1 shows the context attributes and their domain of values.

**Table 4-1: Example of context attributes and domain of value**

| Context Attributes | Domain of Value | Description |
|---|---|---|
| Traffic Congestion | L, M, H | L = low congestion<br>M = medium congestion<br>H = high congestion |
| Rain | T, S, M, H, VH, N/A | T = trace (< 0.1 mm.)<br>S = slight rain ( 0.1 – 10.0 mm)<br>M = moderate rain (10.1 - 35.0 mm)<br>H = heavy rain (35.1 – 90.0 mm)<br>VH = very heavy rain ( > 90.1 mm ) |
| Work Day | Yes, No | Yes = days that are not weekends nor public holidays<br>No = Saturday, Sunday and public holidays |
| Day of the Week | D1, D2, …, D7 | D1 = Sun, D2 = Mon, D3 = Tue, D4 = Wed, D5 = Thurs, D6 = Fri, D7 = Sat |
| Time of Day | p1, p2,…, p24 | Time of the day is allotted into 24 period (p1 – p24), 1 hour per 1 period. |
| School Break | Y, N | Y = days during the school break<br>N = days during the school semester |

### 4.2.1 Influential Context Attributes

The contexts that could contribute to traffic congestion estimation range from sensory data, weather data and known facts (such as school break information, day, time and public holidays) to road and building construction. Even though many and varied context factors could be used to estimate traffic congestion, we need to select those factors with the highest correlation between the context attributes and the observed road segment; anything less may lead to inaccurate results. To reduce the number of calculations and at the same time improve accuracy, the contexts must be filtered to those contexts that impact the traffic congestion of the roads under observation. One important factor in achieving accurate results is the perceived relationship between the road to be observed and the context used. Yin, Junli, and Huapu [124] suggest several root causes of congestion: bottlenecks, traffic incidents, bad weather, work zones, poor signal timing and special events. These causes of congestion can be considered influential contexts along with other related contexts in the environment. Possible influential contexts could also be, but are not limited to, data from traffic sensors, the date and time, disasters such as fire, flood and earthquakes, road and building construction, demonstrations, public holidays, the traffic of connected roads, mobile

users' profiles, cellular information, mobile data initiated from mobile devices, street layout, vehicle type, and traffic reports from mobile user volunteers. In addition, social networking websites such as Facebook and Twitter now provide real time access to "who", "what", and "when" and geolocation now provides the "where" in real time; thus data contributed from these websites could be useful contexts for road traffic estimation. People in social networks could contribute their traffic data both intentionally and unintentionally and sometimes include the temporal and location contexts.

In addition to translating the acquirable contexts into the right format, the context attribute extractor module also filters out irrelevant context attributes. For example, in Bangkok, only rain data in a weather log is taken into account because of its impact on traffic congestion. Furthermore, only the traffic congestion data, time stamp and link ID, which influence traffic flow, are chosen from traffic logs.

The influential context attributes considered in our initial framework came from our literature review and from recommendations by experts who are researchers in ITS at National Electronics and Computer Technology Center of Thailand (NECTEC).

## *4.3 Compensating for Missing Sensory Traffic Data*

To compensate for missing sensory data, we could apply well-known methods such as replacing the missing data with a probable value or using a data mining technique. The principle of each of these approaches is explained in the next section. However, these two approaches have limitations that the CATE framework overcomes. An evaluation of the performance of all three approaches can be found in Chapter 5.

### 4.3.1 Using Mode Value Approach

Finding missing sensory traffic data is similar to finding missing values in data mining. Several classification algorithms that work with missing data are usually based on filling in a missing value with the most probable value [95].

Replacing the missing value with the mode or an average value is a common static method to solve missing data problems. The missing data or missing value is replaced with the mode (when using nominal data) or with the average (when using numeric

data). The mode value used to replace missing data is calculated from historic data [94]. Since the data we use for experiment is nominal, when the program detects that the sensory data to be reported of a particular road segment is missing, the system reports the mode value as the inferred traffic congestion degree to users. The mode is obtained from historical data. The flow of the mode approach is illustrated in Figure 4-4.

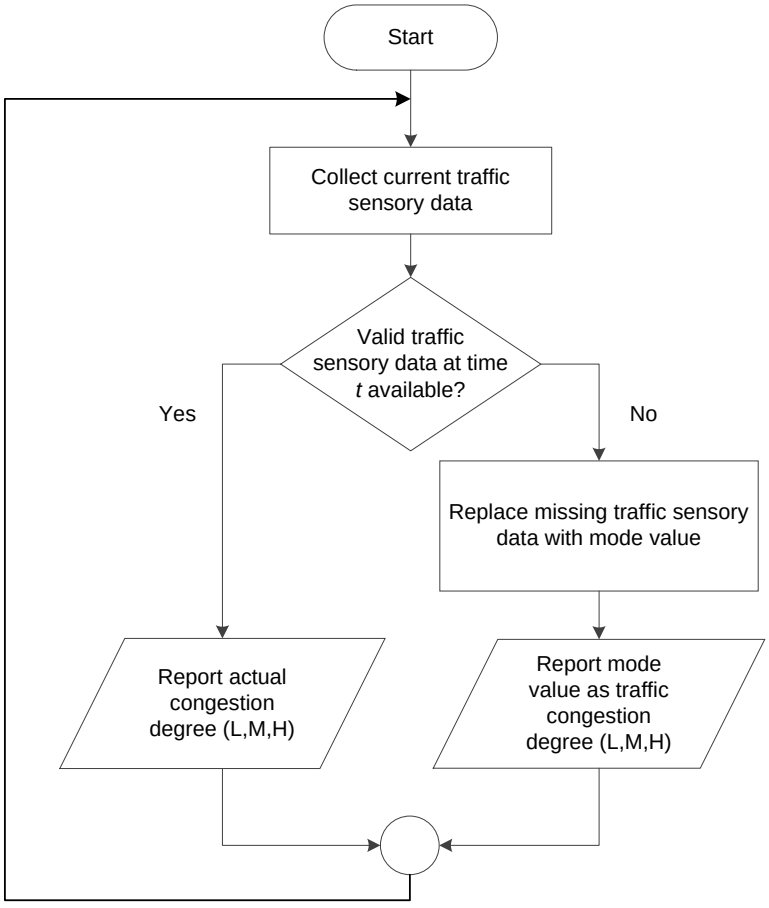



**Figure 4-4: Flow of using mode approach**

Using a mode value to replace missing sensory data is simple and intuitive, but it can lead to errors. Errors are likely to be greater when the actual traffic data is *H*, as the system always infers the missing traffic data based on the mode value *L*. The difference between an *H* and *L* traffic congestion degree is large enough to be

meaningful because it can lead to road users making incorrect route decisions. This issue is discussed further in the following chapters.

As discussed, relying on a data mining technique that uses only sensory traffic data has limitations. An alternative is to use a data mining technique in combination with available contexts at the time that the traffic congestion degree needs to be inferred (which is the time that the actual sensory traffic data is missing). This approach is discussed in the next two sections.

### 4.3.2 The Single Model Approach

As stated above, one way to compensate for missing sensory data is to use a data mining technique. We can build a model from historical data to infer the traffic congestion degree when the sensory traffic data is not available. We call this approach a single model approach. One popular data mining methodology is the CRISP-DM methodology [91] which is a codification of the data mining process (explained in Chapter 2). The single model approach relies on the CRISP-DM methodology. The process of the single model approach comprises two phases: the learning phase and the inference phase, as illustrated in Figure 4-5.

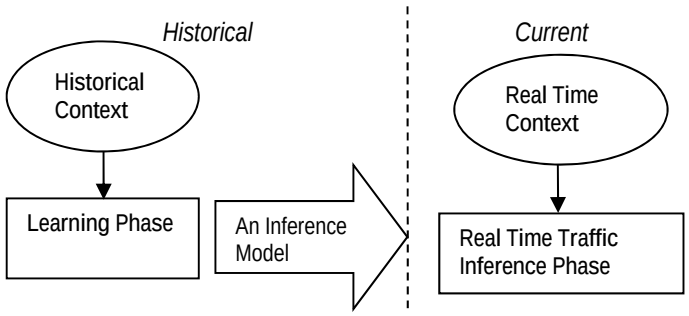



**Figure 4-5: The two main phases of the single model approach**

In the single model approach, only one model (e.g. the decision tree) will be built from all context attributes. The inference model building phase is a learning stage to create a suitable inference model that will be used for inferring the missing sensory traffic data when the system detects that the actual sensory traffic data of a particular road segment is missing at a particular time.

Figure 4-6 describes the learning stage of the single model approach. The historical data is passed through the context attribute extractor that we proposed in section 4.2 Context Attributes and the Context Attribute Extractor. Only influential context attributes will be selected. The influential context attributes that we propose to use are explained in section 4.2.1 Influential Context Attributes. Filtered context attributes then become inputs for the machine learning module to build an appropriate inference model.

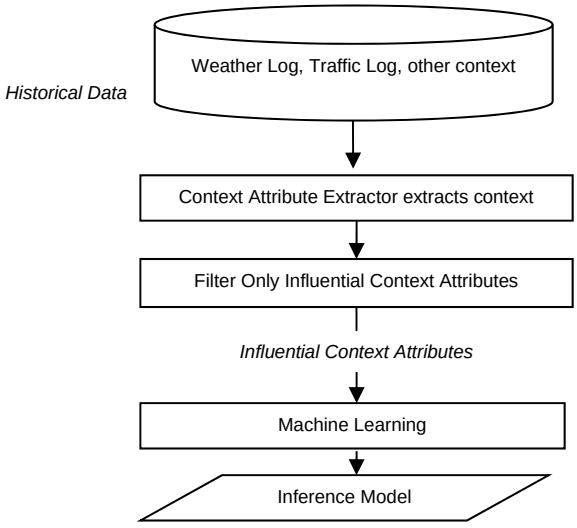



**Figure 4-6: Flow of the learning phase (single model approach)**

A model obtained from the learning phase is used as an inference model in the inference phase. In other words, in this solution we include all the context attributes necessary to build the model.

The flow of the overall process of single model approach is explained in Figure 4-7.

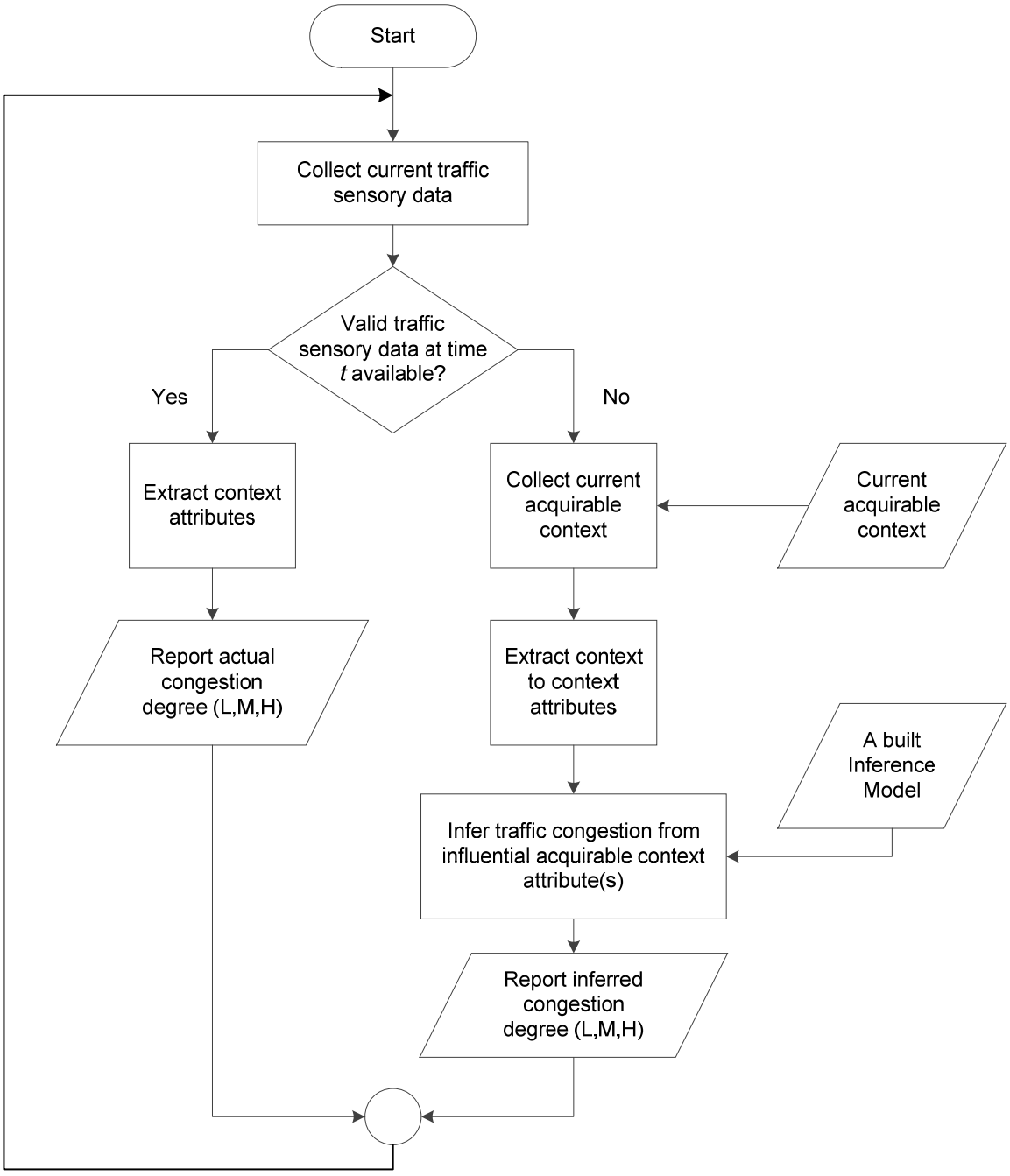



**Figure 4-7: Flow of single model approach**

The process of single model approach starts with learning historical data in order to build an inference model using a selected machine learning algorithm. The built model is evaluated and saved for inferring the traffic condition when necessary. After the model is saved, we deploy the built model to infer the missing traffic congestion degree.

As we propose to use the available context of the environment to infer the traffic congestion degree, when the system detects that the sensory data to be reported of a particular road segment is missing, the system tries to collect the available context of the environment. Then it uses the available influential context attribute(s) as inputs for a pre-built inference model as illustrated in Figure 4-7.

### 4.3.3 The Multiple Models Approach

An issue of using the single model approach for data mining is that accuracy is reduced when the input of the inference model (the influential context attribute value) is not complete. For instance, if the inference model were to be built from six context attributes but during the inferring time only four context attributes were available at the time, there would be two missing attribute values. The missing value estimation is then applied based on the mechanism of each machine learning algorithm selected. Despite this workaround, the missing value will negatively affect the accuracy of the output [95, 125, 126]. Even though methods to deal with missing values have been developed, avoiding missing values in the first place can give better accuracy.

To avoid such negative effects on accuracy, we thus propose to build multiple inference models, with each model being built from different subsets of possible influential context attributes. We call this approach the multiple model approach. For the multiple model approach, if we have $n$ influential context attributes, we will obtain $N_m = \sum_{k=1}^{n} {}^{n}C_k$ models for all possible combinations of influential context attributes where $n$ is the total number of influential context attributes and $k$ is the number of member(s) in each context attribute set.

We propose that the multiple models approach use current available contexts from the environment rather than rely just on sensory data. This contextual data is then used as input for the inference models to calculate the inferred traffic congestion degree in order to compensate for the missing traffic data.

The multiple models approach also has two phases: the learning phase to build the inference models and the inference phase that deploys the inference models to infer the traffic congestion degree (illustrated in Figure 4-5). The flow of the model building phase is similar to Figure 4-6, but is different in the number of inference

models to be built. Now the number of inference models to be built is $N_m$ models. After the models are built, they are saved for future use.

In deployment, if sensory traffic data is found to be unavailable, the program starts searching for a suitable inference model that matches the current available context attributes. The selected model is then used to infer the inferred traffic congestion degree. This technique avoids incomplete input data to the inference model, and hence avoids negatively affecting accuracy.

The artefact that we propose in this research is also based on the CRISP-DM methodology. However, for model building, we propose the multiple models approach to avoid negatively affecting accuracy caused by missing input attribute values for inference models as stated above. In addition, we combine the relearning mechanism to improve the robustness of sparse historical data. The detail of our proposed framework can be found in the next section.

## *4.4 A Context-Aware Traffic Congestion Estimation Framework to Overcome Missing Sensory Data: the CATE Framework*

The proposed framework, CATE framework, is the multiple models approach which consists of two important phases, the *inference model building phase* and the *real time traffic inference phase*. The inputs for both phases are the influential contexts that affect the roads being monitored. Initially, the influential contexts were selected based on the literature review and the recommendations of experts who are researchers in ITS at the National Electronics and Computer Technology Center of Thailand (NECTEC).

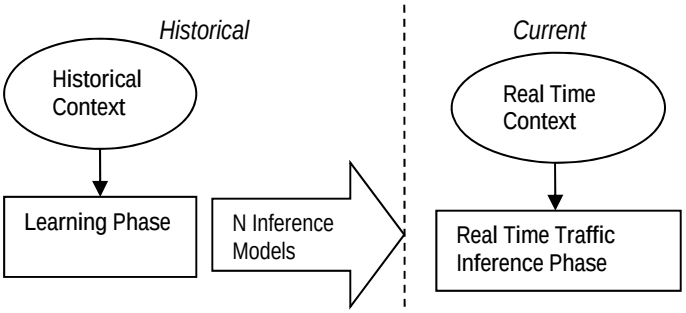



**Figure 4-8: The two main phases of the CATE framework**

The overall process of our CATE framework is demonstrated in Figure 4-8 and Figure 4-9. The process starts by building different inference models suitable for different sets of context attributes in historical time. Each inference model is created for each set of real time acquirable context attributes; each inference model matches each set of real time acquirable context attributes. A matching inference model is chosen in current time (the right hand side in Figure 4-8) depending on the available context at the time.

In the learning phase, we build specific inference models for different sets of real time acquirable context attributes. This avoids the issue of missing data that negatively affects accuracy as explained in Section 4.3.3 The Multiple Models Approach. In addition, our approach aims to create a best fit between the selected inference model and the real time input in order to create a more accurate result. Even though building separate inference models for different sets of context attributes requires more processing and hence a lengthier learning phase, the real time inference phase, which we consider of greater importance, remains unaffected. Because we value minimal processing time in the real time inference phase, we must reduce the time required to calculate the inferred traffic congestion degree as much as possible. In an implementation, each built inference model for a different context attribute set is treated as a case selection. The program will first check the acquirable influential context attributes at a particular time and select the case (each case lead to each matching model) that matches the set of available context attributes.

**Table 4-2: Context attribute sets and their inference models**

| Context attribute set i | Member(s) in context attribute set | Inference Model (IM)*m* |
|---|---|---|
| CtxSet1 | $CtxAtrb_1$ | $IM_1$ |
| CtxSet2 | $CtxAtrb_2$ | $IM_2$ |
| CtxSet3 | $CtxAtrb_3$ | $IM_3$ |
| CtxSet4 | $CtxAtrb_4$ | $IM_4$ |
| CtxSet5 | $CtxAtrb_1$, $CtxAtrb_2$ | $IM_5$ |
| ⋮ | ⋮ | ⋮ |
| CtxSet i | $CtxAtrb_1$, $CtxAtrb_2$, … $CtxAtrb_j$ | $IM_m$ |

From Table 4-2 the members of a context attribute set are the context attributes used to build the different inference models. The context attribute set *i* (CtxSet i) may comprise $CtxAtrb_1$, $CtxAtrb_2$,..., $CtxAtrb_j$ context attributes. The inference model for such a context attribute set is created from a selected machine learning algorithm by using $CtxAtrb_1$, $CtxAtrb_2$,..., $CtxAtrb_j$. In learning phase, the suitable inference model is created for the set of acquirable context attribute(s). For instance, if the acquirable context attributes in the inference phase are $CtxAtrb_1$ *and* $CtxAtrb_2$, the inference model $IM_5$ is selected. (Further explanation of context attribute sets and inference model building can be found in Section 4.4.1.)

Once a suitable model is selected, the real time acquirable context attributes are used as input for the selected model to generate an *inferred traffic congestion degree*. The *inferred traffic congestion degree* is used to compensate for the missing sensory data in the traffic dissemination system. Details of the real time traffic inference process are illustrated in Section 4.4.2 Inference Phase.

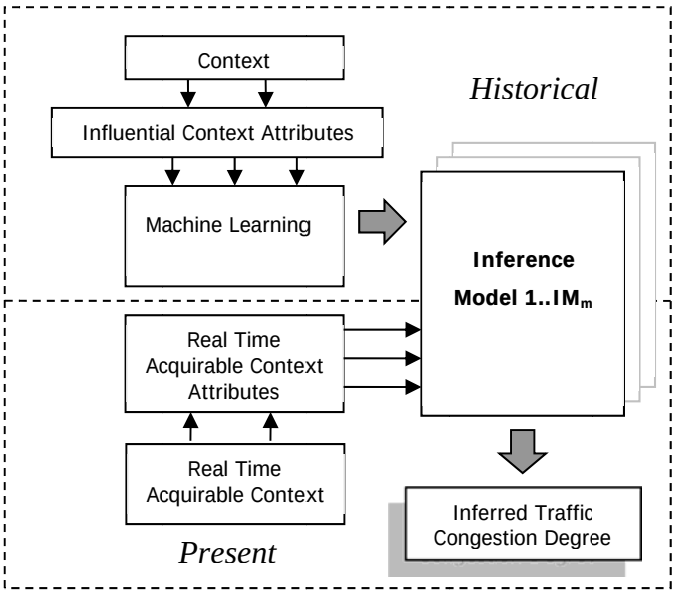



**Figure 4-9: Overall process for computing the traffic congestion degree of sensorless roads using the CATE framework**

For the machine learning component in Figure 4-9, a variety of machine learning algorithms can be applied. At the preliminary stage of our research, we initially tested these various machine learning algorithms using GUI of WEKA [94], a well-known data mining workbench. This testing was undertaken in order to choose a suitable machine learning algorithm that required less model building time, provided high accuracy results and was suitable for our data. We used an Intel Core i7-4500U 1.80 GHz computer with 8GB of memory to conduct this test by using all historical data with six context attributes for learning and building a model based on different machine learning algorithms. The model building time was recorded for each algorithm. After the models were built, we evaluated the obtained models with 10-fold cross validation and recorded the accuracy. A comparison of the model building time and accuracy that resulted from using the different machine learning algorithms when applying the six context attributes for learning is presented in Table 4-3.

The testing revealed that the J48 machine learning algorithm was suitable for our proposed framework. (In the implementation, however, we develop our program using the Groovy programming language and use Weka classes within the program. Groovy is an agile and dynamic language for the Java Virtual Machine that builds upon the

strengths of Java but has additional power features inspired by languages such as Python, Ruby and Smalltalk [127].)

The machine learning algorithms tested were JRip, J48, MultilayerPerceptron, and the NaiveBayes machine learning algorithm. JRip, a RIPPER algorithm, is a rule based classifier. J48 is the C4.5 decision tree learner. MultilayerPerceptron is the back propagation neural network classifier while the NaiveBayes is the standard probabilistic Naive Bayes classifier [94].

**Table 4-3: Comparison of machine learning algorithms when building one model with six context attributes**

| Segment ID to be Inferred | 1206 | | 2613 | | 2718 | |
|---|---|---|---|---|---|---|
| Machine Learning Algorithm | Accuracy | Model Building Time (Sec.) | Accuracy | Model Building Time (Sec.) | Accuracy | Model Building Time (Sec.) |
| JRip | 90.2766 | 153 | 83.0184 | 788 | 99.0155 | 12 |
| J48 | 91.439 | 2 | 86.1904 | 3 | 99.0015 | 1 |
| MultilayerPerceptron | 91.5357 | 4981 | 86.0453 | 4870 | 99.0171 | 5244 |
| NaiveBayes | 87.2373 | 1 | 83.034 | 1 | 98.8532 | 1 |

MultilayerPerceptron is a back propagation neural network classifier with one or more layers between the input and the output layer. In its most basic form, it comprises three layers: the input layer, hidden layer and the output layer. Each neuron in each layer is connected to every neuron in the adjacent layers. The training vectors are connected to the input layer and are further processed by the hidden and output layers [128]. The results in the table indicate that MultilayerPerceptron has a longer learning process than the other algorithms. However, even though MultilayerPerceptron can produce a slightly higher accuracy rate, this gain does not justify the longer model building time. In addition, MultilayerPerceptron does not produce a descriptive model that can be understood by humans.

The NaiveBayes algorithm is a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions between the features [129]. The Bayesian Network is a directed acyclic graph. The assumption made in

BayesNet is that all attributes are nominal and that no values are missing [128]. Although applying NaiveBayes to machine learning required the least model building time of the algorithms, it also had the lowest accuracy rate. The NaïveBayes algorithm was thus not suitable for our proposed framework.

JRip is a rule based RIPPER algorithm designed for fast and effective rule induction. The RIPPER algorithm is a rule based machine learning algorithm that is a class based ordering inductive rule learner. JRip is a learner that builds a set of rules that identifies classes while minimizing errors. The RIPPER algorithm can produce a highly expressive and descriptive model. Its results are thus easy to generate and interpret, as it classifies new instances rapidly yet gives a good performance.

J48 is the C4.5 machine learning algorithm, which is a statistical based decision tree algorithm. It is simple and quick and achieves reliable results [103, 105]. Using J48 as a machine learning algorithm required less model building time and gave very high accuracy (shown in Table 4-3). It is also well suited to our nominal data sets.

Even though the class based ordering RIPPER algorithm, JRip, can quickly produce a descriptive model, giving quite high accuracy and rendering it suitable for our nominal data, using J48 for the machine learning yielded higher accuracy rates and required smaller model building time than JRip. These differences would not be important if only one road segment was under consideration. However, considering the total calculation time that would be required for over 5,241 road segments in Bangkok [130] , the reduced model building time of J48 make it more suitable for our purposes than JRip.

As a consequence of these tests, we choose the J48 algorithm [103], a statistical based data mining approach, for our machine learning in order to build inference models. J48 is simple, interpretable and fast. It also achieves reliable results. Furthermore, it is well suited to our nominal data sets.

### 4.4.1 The Inference Model Building Phase (Learning Phase)

The inference model building phase is a learning stage to create a suitable inference model for each set of context attributes. One model matches one set of real time

acquirable context attributes. The inference models obtained after the learning process are used for real time traffic congestion inference.

Figure 4-10 describes the learning stage. The historical contexts are passed through the context attribute extractor which we proposed in section 4.2 Context Attributes and the Context Attribute Extractor. Only influential context attribute will be selected. The influential context attributes that we propose are explained in section 4.2.1 Influential Context Attributes.

Filtered context attributes become input for the machine learning module to generate appropriate $N_m$ inference models where $N_m$ is the number of sets of context attributes. Each set is composed of possible combinations of acquirable context attributes at the current time. For example, inference model 2 is created for the acquirable context attribute set 2, which is composed of only the "rain" context attribute. The model building process is repeated $N_m$ times for all context attribute sets.

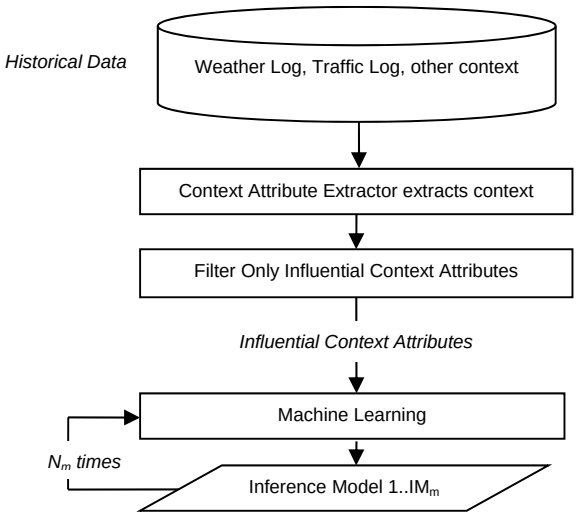



**Figure 4-10: Flow of the learning phase**

As we stated in Section 4.3.3, if we have $n$ influential context attributes, we will obtain $N_m = \sum_{k=1}^{n} {}^{n}C_k$ models for all possible combinations of context attributes where $n$ is the total number of influential context attributes and $k$ is the number of member(s) in each context attribute set.

The input for model building is influential historical contexts, which can be sensory traffic data of connected roads, weather logs, known facts or any available context that has influence on the traffic of the observed road.

The CATE system is self-learning. The relearning task that rebuilds inference models can be repeated at user defined time intervals or whenever new historical data is received. However, determining the optimum time period for relearning is left for future research.

### 4.4.2 Inference Phase

The lower half of Figure 4-9 represents a context aware real time traffic inference stage. The inference models obtained from the learning stage, coupled with real time acquirable context attributes, are used for reasoning to produce real time traffic congestion degrees to compensate for missing sensory data. The brief algorithm of the real time inference process is described in Figure 4-11.

```
Procedure InferCongDegree
/*CtxAtrb is a set of current available context attributes */
Begin
/* Collect available current context */
AllContext := current available context

/* Extract context attribute by function "AtrbExtracter" */
CtxAtrb := AtrbExtracter (AllContext)

/* Filter only influenced context attributes by function "FilterInfluencedCtx" */
CtxSet := FilterInfluencedCtx(CtxAtrb)
        If CtxSet match attribute in          InferenceModel1
        then     Call InferenceModel1 (CtxSet,InferDegree)
                 .....
                 .....
        Else If CtxSet match attribute in     InferenceModelN
        Then Call InferenceModelN (CtxSet,InferDegree)
        End if
/*output the inferred congestion degree*/
output InferredDegree
End
```

**Figure 4-11: Our algorithm for traffic congestion degree inference**

At a particular time, the algorithm tries to collect as much volume of influential context as it can. The collected influential contexts are then converted to context attributes before being fed as inputs to a chosen inference model. The framework then decides the inference model that matches the acquirable context attribute set at real time. The inferred traffic congestion degree is then generated by the selected inference model and reported. The output of our system, which is the inferred traffic congestion degree, can also become the input for estimating other information such as travel time and carbon footprint. However, the inferred traffic congestion degree should be discarded for future re-learning to adjust the new inference models at specific time period (perhaps at every month) to avoid inappropriate learning yielding inaccurate result.

### 4.4.3 An Adaptive Algorithm for Context Aware Traffic Congestion Estimation to Overcome Missing Sensory Data

We propose both an adaptive traffic congestion estimation system architecture for traffic information dissemination as well as a novel traffic condition estimation algorithm that overcomes temporarily missing sensory data. In the case of uncertainty of roadside sensors [121] and intermittently available mobile sensors [122] we can use the proposed algorithm shown in Figure 4-12.

From Figure 4-12, when the sensory traffic data is available, the framework reports actual traffic conditions based on actual data from traffic sensors. However, if the system detects that traffic data is missing, it then adaptively infers the traffic congestion degree from discoverable external contexts using the pre-created inference models. It can thus be seen that our proposed framework can handle uncertainty through this adaptation capability. The framework can adapt and choose a suitable inference model based on the list of available contexts at run time and can automatically retrain itself by relearning process.

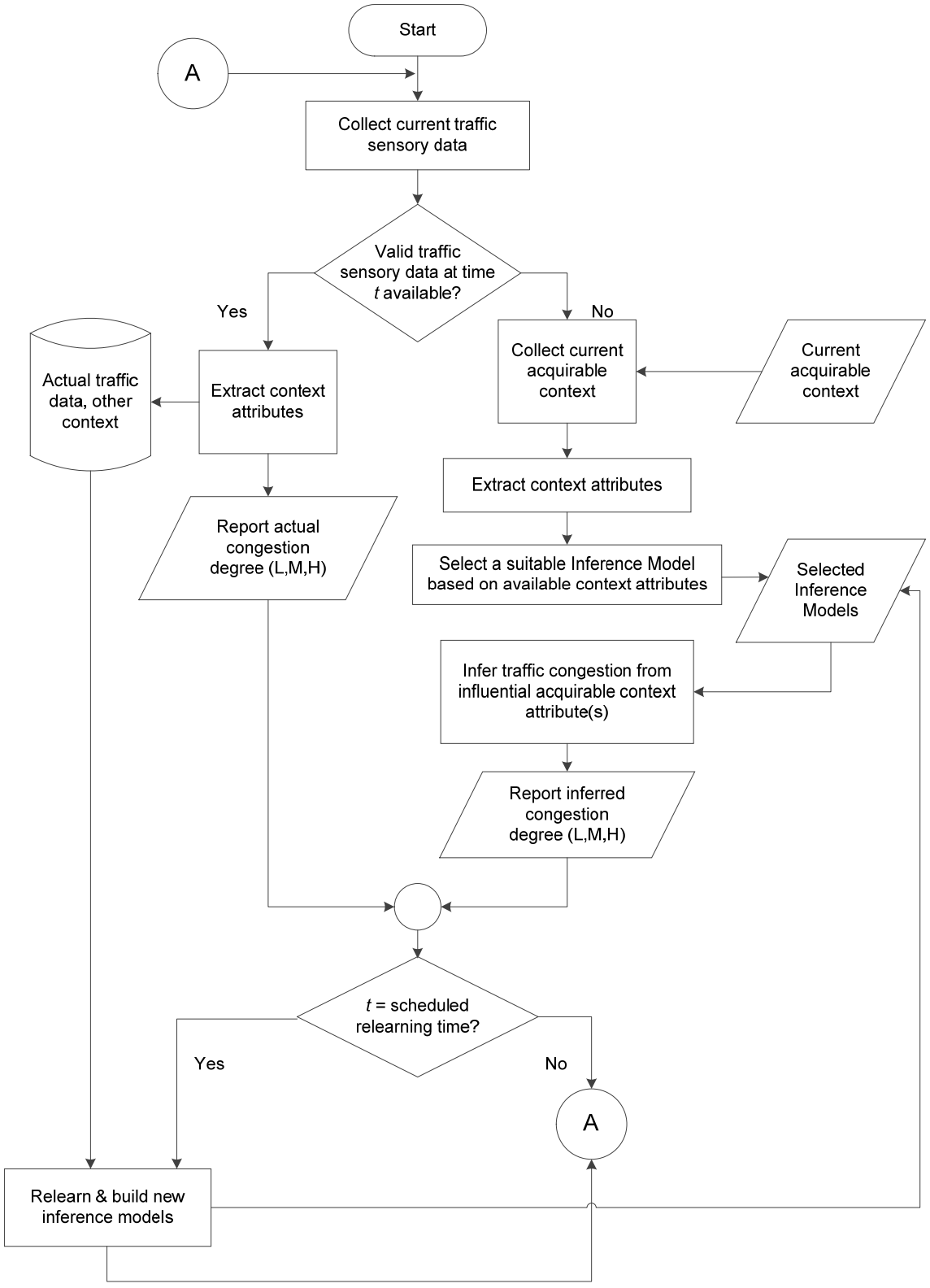



**Figure 4-12: Flow chart of the proposed adaptive context aware traffic congestion estimation system to overcome missing sensory data**

The proposed framework also keeps actual traffic data logs (along with other contexts) for future retraining purposes. Nevertheless, the inferred values are not included in the retraining process in order to avoid skewed results in the future. As previously discussed, the system is set to automatically relearn at a specific time interval (for example, every month) as specified by the program designer. Higher priority is given to more recent historical data.

The relearning and rebuilding process is designed to be robust even under conditions involving small amounts of historical data, such as one month. As time passes, the CATE system keeps new incoming actual data for relearning in the future so that the historical data used for relearning and building new models is constantly growing. The system will thus improve itself automatically as more actual data is collected. In addition, the system can adapt to changes that will occur in contextual factors over time (for example, changes in traffic flow management, the number of vehicles, or buildings in the area (such as a new school or shopping centre). As the pattern and relation of contextual factors change over time, the inference models will adapt to these changes and constantly improve.

Preliminarily, the interval between each invocation of the relearning process is fixed to approximately one month in our experiment. However, further investigation to determine the optimal interval is left for future research because relearning too often may cause high system load and thus performance degradation. This is also subject to the amount of data reflected by the time interval for the relearning process because older data may affect accuracy while larger sizes of data affect relearning time.

It is noticeable that the difference between the single model approach and the CATE framework approach is that the single model approach has only one inference model built while the CATE framework builds multiple inference models that correspond to specific available influential context. The CATE framework thus bypasses the issue of missing values that might yield less accurate results. In addition, the CATE framework has the ability to relearn and adapt to changing environmental conditions at predefined time periods.

In CATE framework, each road segment has its own software module called RoadLink. The class diagram of RoadLink is depicted in Figure 4-13.

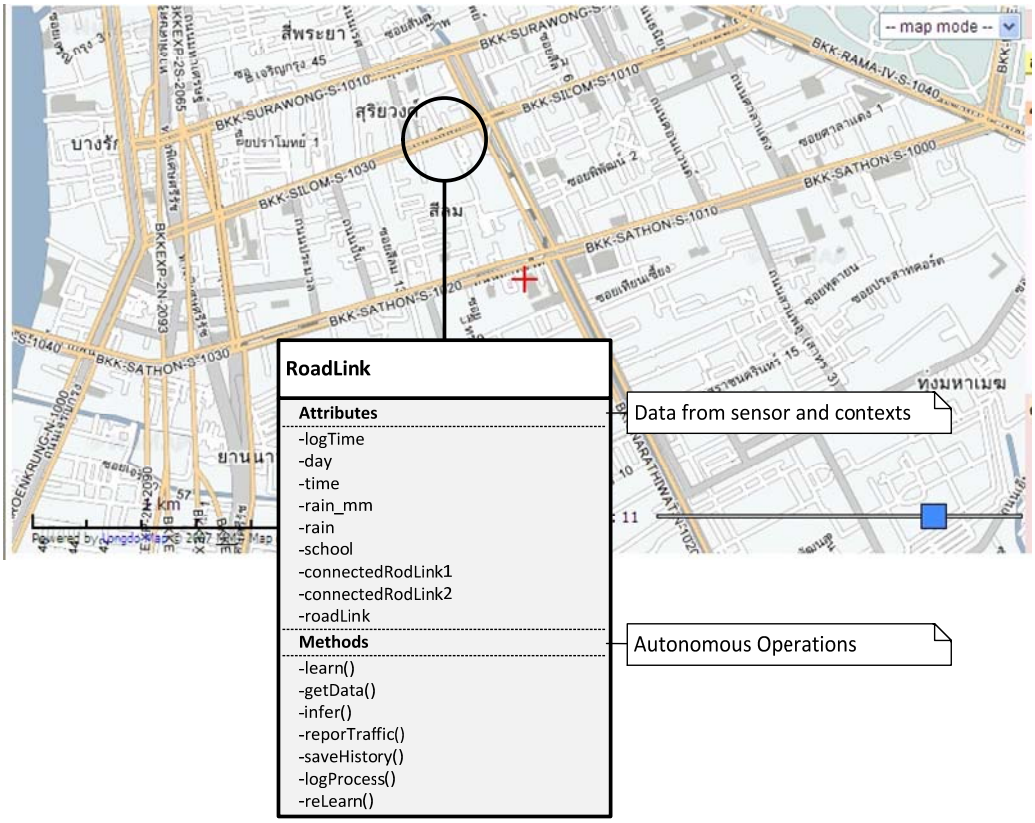



**Figure 4-13: Context Aware RoadLink Class**

For the case of sensorless minor roads (as described in Section 4.1.3) the algorithm comprises the model building phase and the real time inference phase, similar to the other two cases as described in Figure 4-9, but with a slight difference. Even though our approach does not require the sensory traffic data of minor roads at deploy time, the minor roads' historical traffic data for model building in the learning phase is still required. However for minor roads lacking sensors, we can collect traffic data for model building from alternative means. Traffic data can also be collected [123]:

- by residents who live in the area, or volunteers;
- from mobile users or pedestrians who pass that road and send traffic data to the server;
- from traffic mobile sensors that pass the small road at specific times;

- from temporarily installed cost effective sensors such as pressure hose sensors [131] or by using mobile sensors dedicated to collect traffic data in a particular small road for a period of time;
- from temporarily installed cameras at houses or buildings located on the small road (which may be moved to other locations after sufficient traffic data is obtained); and
- from traffic data shared in the driving community through mobile devices with social mobile applications installed such as Waze [132].

These collected data can be used to learn and generate appropriate inference models.

Even though applying our proposed framework to minor roads lacking sensor infrastructure may give a less accurate output than traffic congestion information obtained from dedicated traffic sensors, it can still be considered an alternative method. Nonetheless, we do not recommend obtaining the inferred congestion degree by this method where it will be used as an input to infer traffic congestion for other road segments as it may give a less accurate outcome.

## *4.5 Chapter Summary*

In typical traffic dissemination systems, traffic data is provided from traffic sensors. However, the sensory data may be absent at particular times. Losing sensory data can render a traffic report incomplete and reduce its usefulness. In this chapter, we presented a framework for context aware traffic congestion estimation to compensate for missing sensory data. We addressed the problem of missing data by providing an approach that approximates the traffic congestion degree by utilizing available real time acquirable context rather than relying on sensory traffic data.

Three problematic scenarios that our framework can address were presented in this chapter: the uncertainty of roadside sensors; intermittently available mobile sensors; and sensorless roads. The third situation typically occurs on minor roads where the investment in sensor infrastructure is limited by the capital budget. Traffic information for minor roads is thus inadequate or frequently non-existent.

In addition to the CATE framework, in this chapter we also discussed two approaches (and their limitations) that can be used to compensate for the missing sensory traffic data: the mode approach and single model approach. The comparison of the performance of each approach can be found in Chapter 5.

Our approach is cost efficient, expedient to implement, independent of sensor type and flexible. Furthermore, the proposed framework supports the concept of "anytime", "anywhere" and "anything" for ubiquitous ITS [77]. It also achieves the availability, transparency, seamlessness and awareness goals of ubiquitous services.

The implementation to evaluate the CATE framework is presented in the next chapter. However, we already know that the preliminary framework proposed in this chapter can be improved. To this end, we conducted further investigation into influential contexts to determine the most appropriate constituents of a context data set. The process of this investigation and an evaluation of the resulting improved framework are presented in Chapter 6.

# Chapter 5 Evaluation of the CATE Framework

In order for a traffic report service to obtain continuous traffic data, sensors must send sensory data to the system at pre-determined time periods. However, in some time periods, sensory data may not be generated nor collected, and, consequently, useful traffic information may not be available to drivers. Our approach can alleviate this restriction by utilizing real time discoverable context instead of relying just on sensory data. The CATE framework proposed in the previous chapter is designed to improve the ability of existing traffic report service in ITS that may suffer from missing sensory data.

An implementation of our CATE framework is presented in this chapter for performance evaluation. We simulate a scenario equivalent to a real situation to evaluate the performance. We also compare the performance of the CATE framework to two other potential solutions: the mode approach and the single model approach. We evaluate performance by measuring both the build time and the accuracy of the results obtained from inference process.

## *5.1 Data Collection and Preparation*

### 5.1.1 Data Sources

The data sets used for our evaluation are the traffic data logs of roads in Bangkok. Information based on these traffic logs is disseminated through the media and through devices such as traffic information websites, electronic boards installed at intersections in Bangkok and traffic report applications on mobile phones. This data is administered by the National Electronics and Computer Technology Center of Thailand (NECTEC) and is updated every five minutes on a daily basis. Figure 5-1, Figure 5-2 and Figure 5-3 respectively show examples of a website, electronic boards at intersections and mobile phone applications that display traffic congestion

information of road segments. We use the same data that have been used to produce these announcements to evaluate our framework. These traffic data are produced by the Bangkok Metropolitan Administration (BMA), the Expressway Authority of Thailand (EXAT) and NECTEC's CCTV.

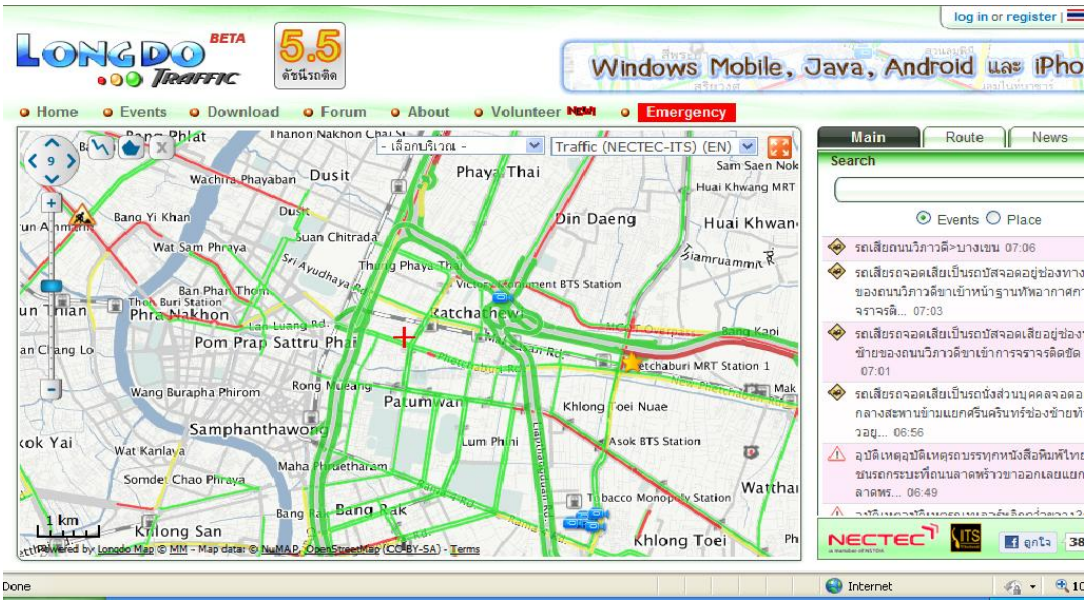



**Figure 5-1: Screenshot of a traffic information dissemination website**


**Figure 5-2: An electronic board showing the traffic congestion degree of road segments at each intersection in Bangkok**


**Figure 5-3: Screenshots of traffic report applications on mobile phones**

The traffic data logs we obtained for use in our experiment are text files in the format shown in Figure 5-4 .

*Date time stamp; location; link id; traffic congestion degree (L, M, H)*

**Figure 5-4: The format of a traffic data log**

The *date time stamp* notes the date and time at which the traffic data was collected.

*Location* notes in which road segment the data was collected. A road is divided into many road segments. A road segment is from one intersection to an adjacent intersection on the same road. Each road segment in Bangkok is given a *link id* of four digits.

The traffic congestion degree in the traffic log is represented as *L* for low, *M* for medium and *H* for high traffic congestion degrees.

An extract of a traffic log example is shown in Figure 5-5. The traffic congestion degree is recorded in the traffic log every 5 minutes.

```
2008-06-29 19:51:28;1214743888;708;L
2008-06-29 19:51:28;1214743888;709;M
2008-06-29 19:51:28;1214743888;710;M
2008-06-29 19:51:28;1214743888;711;M
2008-06-29 19:51:28;1214743888;712;L
2008-06-29 19:51:28;1214743888;713;H
2008-06-29 19:51:28;1214743888;803;L
2008-06-29 19:51:28;1214743888;808;L
2008-06-29 19:51:28;1214743888;811;M
2008-06-29 19:51:28;1214743888;813;M
2008-06-29 19:51:28;1214743888;814;H
2008-06-29 19:51:28;1214743888;904;L
2008-06-29 19:51:28;1214743888;905;L
2008-06-29 19:51:28;1214743888;908;L
2008-06-29 19:51:28;1214743888;909;L
2008-06-29 19:51:28;1214743888;913;M
2008-06-29 19:51:28;1214743888;1001;F
2008-06-29 19:51:28;1214743888;1007;L
2008-06-29 19:51:28;1214743888;1014;L
2008-06-29 19:51:28;1214743888;1016;L
2008-06-29 19:51:28;1214743888;1017;M
2008-06-29 19:51:28;1214743888;1019;L
2008-06-29 19:51:28;1214743888;1020;L
2008-06-29 19:51:28;1214743888;1021;L
2008-06-29 19:51:28;1214743888;1022;L
2008-06-29 19:51:28;1214743888;1101;L
2008-06-29 19:51:28;1214743888;1104;L
2008-06-29 19:51:28;1214743888;1107;L
2008-06-29 19:51:28;1214743888;1202;L
2008-06-29 19:51:28;1214743888;1206;L
```

**Figure 5-5: Example of a part of a traffic log.**

Figure 5-6 illustrates road segments and their link id.

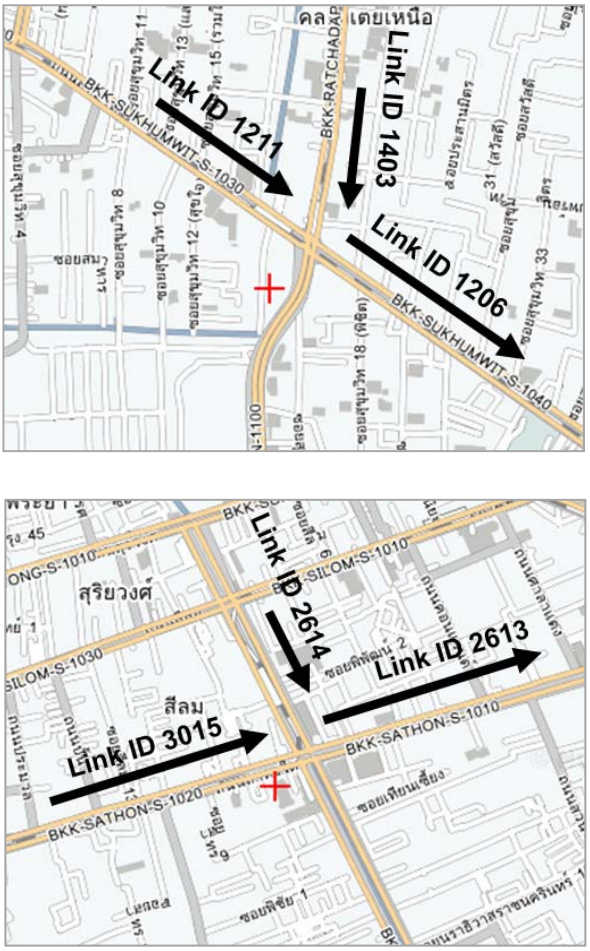
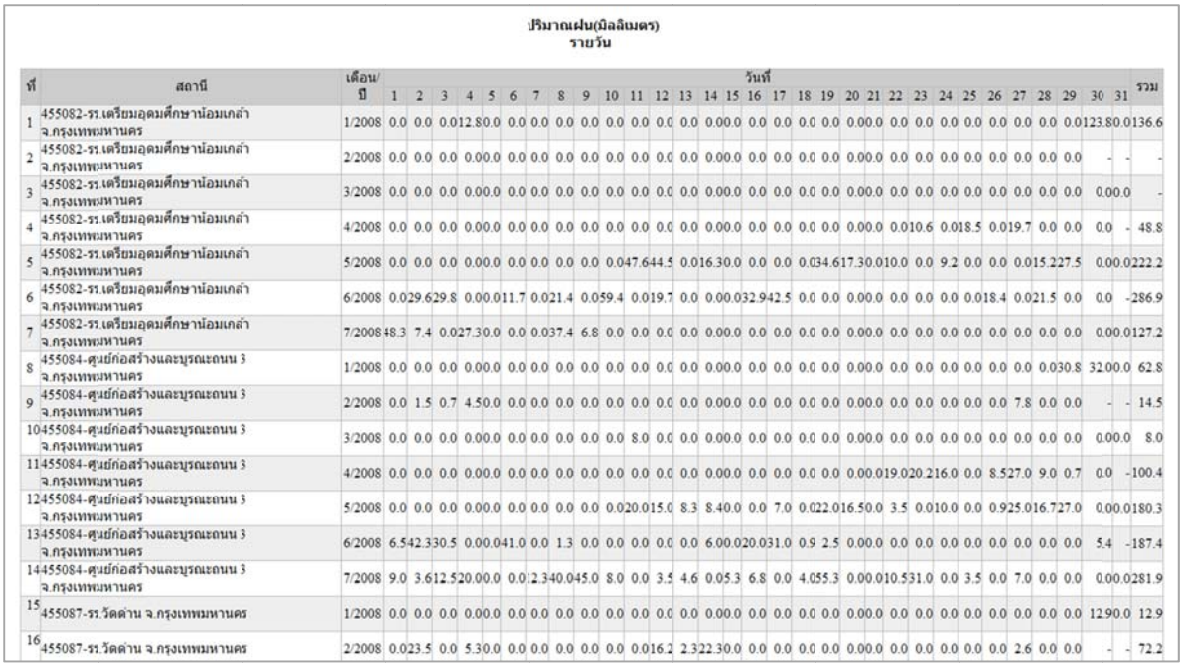



**Figure 5-6: Road segments represented by Link ID**

In addition to the traffic data logs, we also use rain data as one of the context attributes. For our evaluation, the rain data was extracted from the weather log of the Thai Meteorological Department. The log is composed of diverse weather data such as pressure, humidity, temperature, wind direction and visibility. However, only rain data is of interest because of its impact on the traffic conditions in Bangkok. Figure 5-7 shows an extract from the rain volume log in different areas of Bangkok. The rain data we use is taken from the same time periods as the traffic data we use for our evaluation. The rain volume is recorded in millimetres (mm). In other countries, visibility, snow or slush on the road surface may also impact traffic conditions. If relevant, these factors could also be included in the influential context list.



ปริมาณฝน(มิลลิเมตร)
รายวัน

| ที่ | สถานี | เดือน/ปี | วันที่ | | | | | | | | | | | | | | | | | | | | | | | | | | | | | รวม |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | |
| 1 | 455082-รร.เตรียมอุดมศึกษาน้อมเกล้า จ.กรุงเทพมหานคร | 1/2008 | 0.0 | 0.0 | 0.0 | 12.8 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 123.8 | 0.0 | 136.6 |
| 2 | 455082-รร.เตรียมอุดมศึกษาน้อมเกล้า จ.กรุงเทพมหานคร | 2/2008 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | - | - | - |
| 3 | 455082-รร.เตรียมอุดมศึกษาน้อมเกล้า จ.กรุงเทพมหานคร | 3/2008 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | - |
| 4 | 455082-รร.เตรียมอุดมศึกษาน้อมเกล้า จ.กรุงเทพมหานคร | 4/2008 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 10.6 | 0.0 | 18.5 | 0.0 | 19.7 | 0.0 | 0.0 | 0.0 | - | 48.8 |
| 5 | 455082-รร.เตรียมอุดมศึกษาน้อมเกล้า จ.กรุงเทพมหานคร | 5/2008 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 47.6 | 44.5 | 0.0 | 16.3 | 0.0 | 0.0 | 0.0 | 0.0 | 34.6 | 17.3 | 0.0 | 10.0 | 0.0 | 9.2 | 0.0 | 0.0 | 0.0 | 15.2 | 27.5 | 0.0 | 0.0 | 0.0 | 222.2 |
| 6 | 455082-รร.เตรียมอุดมศึกษาน้อมเกล้า จ.กรุงเทพมหานคร | 6/2008 | 0.0 | 29.6 | 29.8 | 0.0 | 0.0 | 11.7 | 0.0 | 21.4 | 0.0 | 59.4 | 0.0 | 19.7 | 0.0 | 0.0 | 0.0 | 32.9 | 42.5 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 18.4 | 0.0 | 21.5 | 0.0 | 0.0 | - | 286.9 |
| 7 | 455082-รร.เตรียมอุดมศึกษาน้อมเกล้า จ.กรุงเทพมหานคร | 7/2008 | 48.3 | 7.4 | 0.0 | 27.3 | 0.0 | 0.0 | 0.0 | 37.4 | 6.8 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 127.2 |
| 8 | 455084-ศูนย์ก่อสร้างและบูรณะถนน 3 จ.กรุงเทพมหานคร | 1/2008 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 30.8 | 32.0 | 0.0 | 62.8 |
| 9 | 455084-ศูนย์ก่อสร้างและบูรณะถนน 3 จ.กรุงเทพมหานคร | 2/2008 | 0.0 | 1.5 | 0.7 | 4.5 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 7.8 | 0.0 | 0.0 | - | - | 14.5 |
| 10 | 455084-ศูนย์ก่อสร้างและบูรณะถนน 3 จ.กรุงเทพมหานคร | 3/2008 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 8.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 8.0 |
| 11 | 455084-ศูนย์ก่อสร้างและบูรณะถนน 3 จ.กรุงเทพมหานคร | 4/2008 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 19.0 | 20.2 | 16.0 | 0.0 | 8.5 | 27.0 | 9.0 | 0.7 | 0.0 | - | 100.4 |
| 12 | 455084-ศูนย์ก่อสร้างและบูรณะถนน 3 จ.กรุงเทพมหานคร | 5/2008 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 20.0 | 15.0 | 8.3 | 8.4 | 0.0 | 0.0 | 7.0 | 0.0 | 22.0 | 16.5 | 0.0 | 3.5 | 0.0 | 10.0 | 0.0 | 0.9 | 25.0 | 16.7 | 27.0 | 0.0 | 0.0 | 180.3 |
| 13 | 455084-ศูนย์ก่อสร้างและบูรณะถนน 3 จ.กรุงเทพมหานคร | 6/2008 | 6.5 | 42.3 | 30.5 | 0.0 | 0.0 | 41.0 | 0.0 | 1.3 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 6.0 | 0.0 | 20.0 | 31.0 | 0.9 | 2.5 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 5.4 | - | 187.4 |
| 14 | 455084-ศูนย์ก่อสร้างและบูรณะถนน 3 จ.กรุงเทพมหานคร | 7/2008 | 9.0 | 3.6 | 12.5 | 20.0 | 0.0 | 0.0 | 2.3 | 40.0 | 45.0 | 8.0 | 0.0 | 3.5 | 4.6 | 0.0 | 5.3 | 6.8 | 0.0 | 4.0 | 55.3 | 0.0 | 0.0 | 10.5 | 31.0 | 0.0 | 3.5 | 0.0 | 7.0 | 0.0 | 0.0 | 0.0 | 0.0 | 281.9 |
| 15 | 455087-รร.วัดด่าน จ.กรุงเทพมหานคร | 1/2008 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 12.9 | 0.0 | 12.9 |
| 16 | 455087-รร.วัดด่าน จ.กรุงเทพมหานคร | 2/2008 | 0.0 | 23.5 | 0.0 | 5.3 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 16.2 | 2.3 | 22.3 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 2.6 | 0.0 | 0.0 | - | - | 72.2 |

**Figure 5-7: parts of rain volume logs (part of weather logs)**

## 5.1.2 Data Preparation and Pre-Processing

Although all of the collected data could be considered contextual data, the context that is relevant and makes traffic state estimation possible depends on each environment. Contextual data could include, for example, traffic incident reports, data from traffic sensors, the weather, the day of the week, the time, any special holiday, user profiles of mobile phone users, cellular information, mobile phone location, mobile data initiated from mobile devices, layout of the streets, building constructions and type of cars. For our proposed framework and the evaluation in this chapter, we chose only the *influential contexts* as shown in Table 5-1 follow the process explained in Section 4.2.1 in Chapter 4.

The influential context must first be converted into an appropriate form - a *context attribute* - before being used by the context attribute extractor. For our evaluation, some of the *numeric data* was converted into a *nominal format*. For instance, the rain data in the weather log was converted from a numerical format (in mm) to *T* (< 0.1 mm), *S* (0.1 – 10.0 mm), *M* (10.1 - 35.0 mm), *H* (35.1 – 90.0 mm) or *VH* (> 90.1 mm). The conversion followed the format of the Thai Meteorological Department. The

converted format and the description of each context attribute are described in Table 5-1.

**Table 5-1: Selected influential context attributes converted to nominal format**

| Influential Context Attribute(s) | Nominal Value | Description |
|---|---|---|
| Traffic congestion | L, M, H | L = low congestion<br>(0.00 ≤ OR ≤ 0.30)<br>M = medium congestion<br>(0.31 ≤ OR ≤ 0.80)<br>H = high congestion<br>(0.81 ≤ OR ≤ 1.00) |
| Rain | T, S, M, H, VH, N/A | T = trace<br>(less than 0.1 mm)<br>S = slight rain<br>(0.1 – 10.0 mm)<br>M = moderate rain<br>(10.1 - 35.0 mm)<br>H = heavy rain<br>(35.1 – 90.0 mm)<br>VH = very heavy rain<br>( 90.1or higher) |
| Day | D1, D2, D3, D4, D5, D6, D7 | Day of the week<br>D1 = Sunday<br>D2 = Monday<br>D3 = Tuesday<br>D4 = Wednesday<br>D5 = Thursday<br>D6 = Friday<br>D7 = Saturday |
| Time period of the day | p1, p2, p3, p3, p5, p6, p7, p8, p9, p10, p11, p12, p13, p14, p15, p16, p17, p18, p19, p20, p21, p22, p23, p24 | Time period in one day, 1 hour per 1 period.<br>p1 = 0000–0100<br>p2 = 0101–0200<br>p3 = 0201–0300<br>p4 = 0301–0400<br>p5 = 0401–0500<br>p6 = 0501–0600<br>p7 = 0601–0700<br>p8 = 0701–0800<br>p9 = 0801–0900<br>p10 = 0901–1000<br>p11 = 1001–1100<br>p12 = 1101–1200<br>p13 = 1201–1300<br>p14 = 1301–1400<br>p15 = 1401–1500<br>p16 = 1501–1600<br>p17 = 1601–1700<br>p18 = 1701–1800<br>p19 = 1801–1900<br>p20 = 1901–2000<br>p21 = 2001–2100<br>p22 = 2101–2200<br>p23 = 2201–2300<br>p24 = 2301–2400 |
| School break | Y, N | Y = days during the school break<br>N = days during the school semester |

The context attributes used in our experiment are ***school break*** (*Y*: day is in the school break period, *N*: day is not in the school break period), ***rain*** (*T*: trace, *S*: slight rain, *M*: moderate rain, *H*: heavy rain, *VH*: very heavy rain), ***day*** (*D1*: Sunday, *D2*: Monday, *D3*: Tuesday, *D4*: Wednesday, *D5*: Thursday, *D6*: Friday, *D7*: Saturday), ***time*** (time of the day allotted into 24 periods *p1-p24*) and ***traffic congestion degree of road segments*** (*L*, *M* and *H*). *L*, *M* and *H* equate to the occupancy ratio (OR) in the traffic engineering discipline [133]. The traffic congestion degree *L* equates to OR values between 0.00 and 0.30 inclusive. The *M* value equates to OR values between 0.31 and 0.80 inclusive. The *H* value equates to OR values between 0.81 and 1.00 inclusive.

After the data sets are extracted into context attributes and converted into suitable forms, the model building in the learning stage can be performed.

For our evaluation, we choose data sets from three road segments with link IDs 1206, 2613 and 2718 respectively (shown in the following figure). These segments were chosen because they are typical of many of the roads in Bangkok. In addition, these three road segments were chosen because of their proximity to the rain sensor installed at the nearby Queen Sirikit National Convention Center. The advantage of the rain sensor's proximity is the accuracy of the rain data for these road segments. However, because this one sensor provides rain data for the whole of the Bangkok area, rain data may not be so accurate for road segments in areas that are far from a place rain sensor is installed.

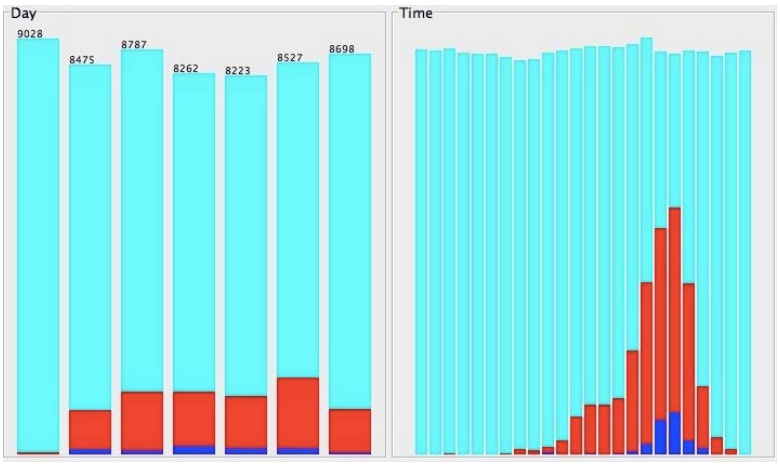



**Figure 5-8: Data visualization of road segment ID 1206**

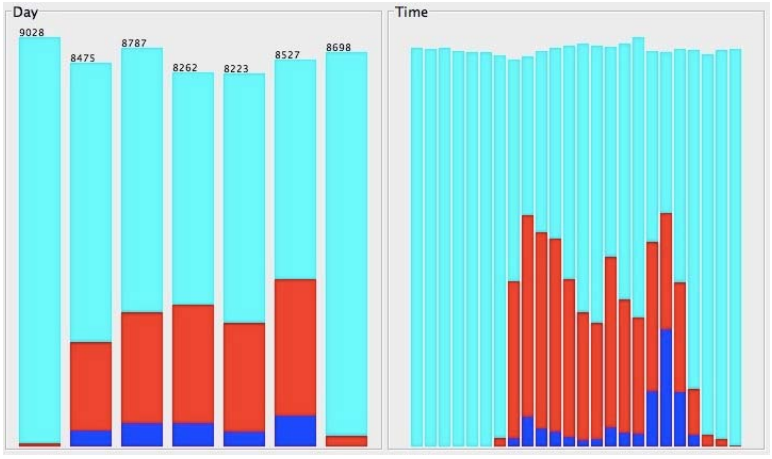



**Figure 5-9: Data visualization of road segment ID 2613**

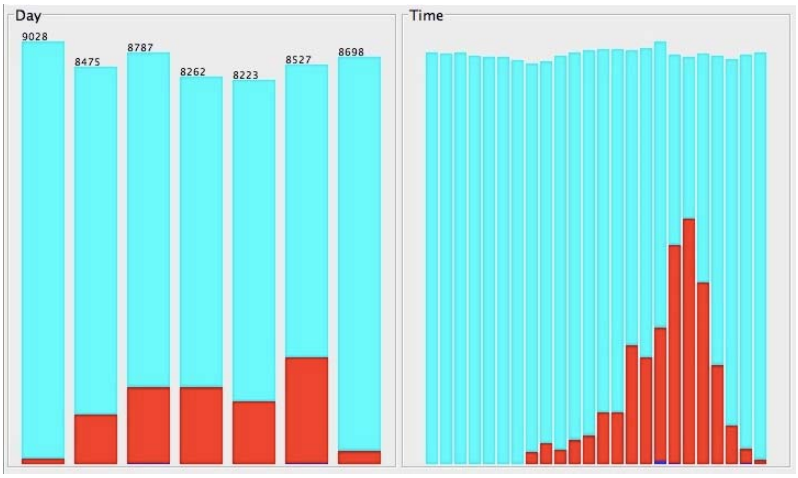



**Figure 5-10: Data visualization of road segment ID 2718**

Figure 5-8, Figure 5-9, and Figure 5-10 show the distribution of the traffic congestion degree for each day of the week and for each time period in each day for the three road segments. This distribution is representative of most road segments in Bangkok.

The light blue color represents the *L* traffic congestion degree, the red represents the *M* traffic congestion degree and the dark blue represents the *H* traffic congestion degree. As explained in Table 5-1, the bars in the "Day" charts start from D1 (Sunday) to D7 (Saturday). The bars in the "Time" charts start from p1 (0000–0100) to p24 (2301–2400).

It is noticeable from the data visualization that the great majority of the traffic congestion degree for all days and all time periods is *L* (light blue). The variation of *L*, *M* and *H* traffic congestion degrees occurs during weekdays (Monday to Friday) from 0600 hours to 2300 hours on each day. We can thus categorize traffic into active and non-active periods. The non-active period is anytime on Saturday (D7) and Sunday (D1) and from 2300 hours to 0600 hours (p24 and p1-p6) on Monday (D2) to Friday (D6) when the traffic condition is *L* almost all of the time. The active period, which involves all three *L*, *M*, and *H* traffic congestion degrees, occurs on weekdays between 0600 hours and 2300 hours. This designation and recognition of active and non-active periods has implications for the evaluation of our results (discussed in Section 5.4).

The road segment 1206 receives the traffic flow from the road segments 1211 and 1403, and thus the traffic congestion degrees of these two road segments are used in the estimation. Similarly, each of the observed road segments 2613 and 2718 receives traffic flow from two connected road segments, whose traffic congestion degrees are also included inference phase as shown in Table 5-2. For simplicity, we rename the context attribute *traffic congestion degree* of each road segment to the *link ID* of the connected road segments.

**Table 5-2: Context attributes used for each observed road segment**

| Observed Road Segment | Context Attributes |
|---|---|
| 1206 | day, time, rain, school, 1211, 1403 |
| 2613 | day, time, rain, school, 2614, 3015 |
| 2718 | day, time, rain, school, 2515, 1401 |

## 5.2 Evaluation Setup

Generally, traffic report services deliver information to users by obtaining traffic data from sensors. The sensory data is processed to yield specific traffic information to be reported to road users. In order for systems to provide continuous traffic data, the sensors must send data to the systems at specified time periods. However, the availability of sensory data may be intermittent and hence cause a disruption to the provision of traffic information. The CATE framework overcomes this limitation by utilizing discoverable context to infer the traffic congestion degree and report to users.

To evaluate the CATE framework, we show the implementation of this approach and two others (the mode and single model approaches) as explained in Section 4.3.1, 4.3.2, and 4.4 in Chapter 4. We compare the performance of all approaches. To achieve this, we simulate a real situation for our evaluation, as explained in the next section.

### 5.2.1 Simulating a Real Situation for Evaluation

To evaluate the performance of the CATE framework, we created a program using the Groovy programming language [127] to simulate the situation and evaluate the performance of each approach using real traffic data as described in Section 5.1.1. As discussed in the previous chapter, Groovy was chosen because of its agility and features. Using Groovy programming language can also make the code shorten. Some Weka classes are used within the program. Examples of selected programming code are presented in Appendix B. In this study, we ran tests using an Intel Core i5 2.90 GHz computer with 8GB of memory.

The general evaluation process involves three steps: simulating a real situation, applying the potential approaches and then analysing the results.

First, to simulate a real situation, the sensory traffic data of an observed road link is assumed to randomly fail (thus creating the necessity to generate inferred data). Other context attributes are also assumed to randomly fail. We simulate the original missing data situation by creating initial data sets that have specific different percentages of missing values.

Second, we then apply, in turn, each of the three approaches discussed in Section 4.3.1, 4.3.2 and 4.4 in Chapter 4. To maintain authenticity, a portion of the context attributes in each column are also assumed to fail. To achieve this, we randomly introduce missing data into the testing data sets in each column. The level of missing data is set at 20%, 40%, 60%, 80% and 100% (discussed further below).

Third, and finally, the performance of the three approaches is analysed in terms of accuracy and the time taken.

Because the CATE framework must have practical application, we place importance on simulating the real environment and including missing data. The following discussion explains in detail how this is achieved.

In an implementation, we use a random function to generate the missing values that represent the initial traffic data for the relevant road segment, equal to a defined failure rate. A 20% failure rate means that in 10,000 records, 2,000 records of the sensory data from a given road segment will be set to NULL. In addition, we also use a random function to generate missing values for other context attributes in each record at a specific failure rate percentage. The 20% failure rate here means that in 10,000 records, 2,000 records in a particular column will be set to NULL.

In the following example that explains how we simulate the data, we assume we want to report the traffic congestion degree for road segment 1206 and that the sensory data failure rate is 20%.


**Figure 5-11: An example of a relevant road segment (segment 1206) and its connected road segments (road segment 1211 and road segment 1403)**

In this example, the influential context attributes that impact road segment 1206 are the actual traffic congestion degrees garnered from sensors of the connected road segments, which are road segments 1211 and 1403 as shown in Figure 5-11. Additional influential context attributes are rain, day, school and time (see Table 5-1 for more explanation). If we set the failure rate to 20%, the columns in Figure 5-12 will have 20% of all data missing, with the exception of the columns for day and time because these can be obtained from the system clock.

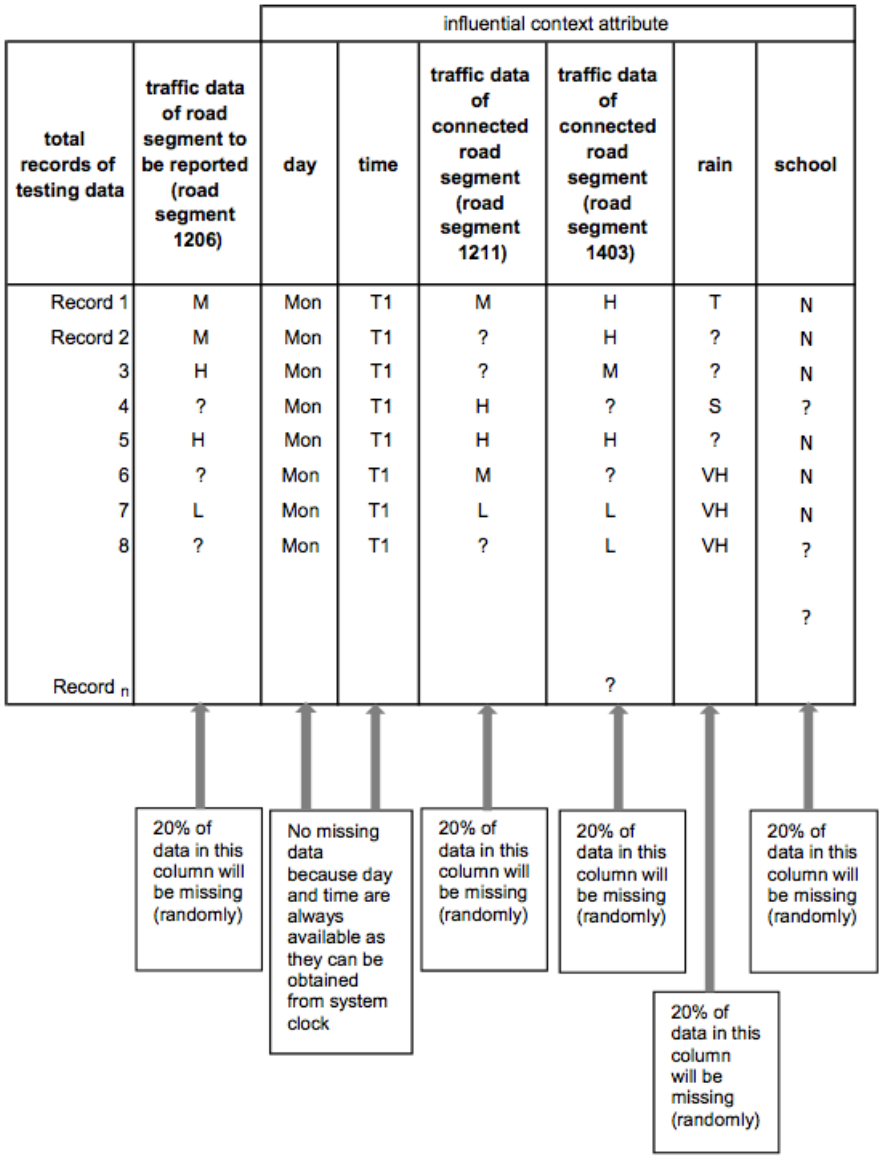



**Figure 5-12: Example of data in records when failure rate is defined at 20%**

From Figure 5-12, each column represents a context attribute. Each row or record represents the context attributes at every five minutes. The "?" symbol indicates that the value of that context attribute in that record is missing. If we consider Record 2, we can see that the sensory traffic data of road segment 1211 is missing and that the rain data is also missing. However, the sensory traffic data of road segment 1206, which is the road segment that we need to report, is not missing. Thus the program can report the actual value to users and does not have to use influential context attributes to infer the traffic congestion degree. On the other hand, if we consider Record 4, the

sensory traffic data of road segment 1206 is missing. In this case, the system will infer the traffic congestion degree of road segment 1206 to compensate for the missing sensory traffic data. The program starts trying to collect possible influential context attribute that are available at that point of inference. The day and time will always be available because they can be obtained from the system clock, but the traffic data of road segment 1403 (the connected road segment) and school data are not available. Thus the only influential context attributes that can be used are day, time, traffic data of road segment 1211 and rain data.

Incorporating random missing data simulates a real situation because there is no guarantee that actual influential context attributes will be permanently available. The traffic sensors of connected roads, or any other sensor (such as a rain sensor) could also fail. In addition, connections may be lost at any time. Nor can school data be guaranteed, as the server that maintains the school data (assuming the data is kept remotely from the CATE system) may not be available.

For the single model and the CATE framework approaches, the system uses influential context attributes only when the sensory traffic data of the relevant road segment (in Figure 5-12, road segment 1206) is missing. However, the mode approach does not require any influential context attributes because whenever the traffic system detects that the sensory traffic data of a road link is missing, the system instantly reports the mode value as the inferred traffic congestion degree to users without any additional calculation.

During the evaluation we repeat this process ten times, record the results and calculate the average degree of accuracy. The reason for repeating the task ten times and calculating an average is that the data for testing differs in each run. This is because we randomly insert NULL values to simulate the missing data. After running the case at a 20% failure rate, we repeat the same process with 40%, 60%, 80% and 100% failure rates and record the results.

## *5.3 Implementation*

To evaluate the CATE framework, we implement it along with two other approaches. The detail of each approach is explained in section 4.3 and 0 in Chapter 4. In this

section, we recapitulate (briefly) the main principles and methods associated with each approach and then describe the implementation details of each approach.

### 5.3.1 The Mode Approach

As explained in Section 4.3.1 in Chapter 4, under the mode approach, the missing value, when using nominal data, is replaced with a mode value. The mode value used to replace the missing data is calculated from historical data [94]. This substitution occurs when the system detects the absence of sensory traffic data in an observed road segment. The mode value, as the compensating traffic congestion degree, is then reported to road users.

The total real data we obtained for conducting experiment is 64,093 records. However, to make it more simplify, we choose to use 60,000 records for evaluation (the same number in three approaches). The data is split into two parts. The first 10,000 records are used for calculating the mode value. The remaining 50,000 records, which are more recent than the first 10,000, are used for testing. A real situation is simulated by incorporating levels of missing data set at 20%, 40%, 60%, 80% and 100% respectively, as explained in section 5.2.1 Simulating a Real Situation, throughout these 50,000 records of testing data.

### 5.3.2 The Single Model Approach

The single model approach follows the general data mining method of building only one inference model from all context attributes to infer the missing sensory data. In this approach, a single decision tree is used as an inference model. In other words, in this solution we include all context attributes to build a model.

Figure 5-13 (below) simplifies the explanation of the implementation of the single model approach. It shows the part of the Groovy programming code [127] that acts as the main method that calls other sub-methods (this is to simplify the code and make it more readable). The code shown here is the core piece of code; the sub-methods can be found in an Appendix B.

```
def process() {

      init()     // learn and build an inference model from all context attributes
                 // then save a built model for inferring later


      // create object for raw testing data
      def dataFile = new File("./ex_data/${name}sensor.and.context.csv")

      // read data record by record (equivalent to real situation that the data come every 5 minutes.)
      dataFile.eachLine {line, number ->

            if(number == 1) return

            getData(line, number)     // extract acquirable context to context attribute(s)
            randomMissing(missRate)    // randomize missing sensory traffic data of observed road and
                                       // other context attributes according to specified missing rate
            infer()    // if the sensory traffic data of observed road is missing
                       // load a built inference model and infer traffic congestion degree
            reportTraffic()    // report actual (sensory traffic data is not missing) or inferred traffic congestion
degree
      }   // loop back to read next record
      summarize()   // generate summary report for experiment and evaluation purpose

  }
```

**Figure 5-13: Extract from the Groovy programming code for the single model approach**

The process starts with the programme learning from historical data in order to build an inference model using the J48 machine learning algorithm. The selection of this algorithm was discussed in Chapter 4. After a model is built, we evaluate the model by using a stratified 10-fold cross validation method to avoid any bias caused by the particular sample chosen for training and testing. The data is divided randomly into ten folds. Each sampled data set represents approximately the same proportions as that of the full data set. Each fold is then held out in turn and the remaining nine folds are used for training. The error rate is calculated against the held out set. Finally, the ten error estimates are averaged to yield an overall error estimate [134].

After the model is evaluated, we use the built model to infer the missing traffic congestion degree. As we explained in Section 4.3.2 of Chapter 4, when the system detects absent sensory traffic data, the system uses values from the available current contexts (for example, rain volume or school break) as the input for the built model to obtain the inferred outcome. If context attributes at the current time (that is, the point of inference) are not available, then a missing value in the decision tree will occur. For

the single model approach, the missing value manipulation is based on the method of a C4.5 decision tree, distribution-based imputation [95, 103, 135].

Like the evaluation of the mode approach, in our evaluation of the single model approach, we use a total of 60,000 records. For the single model approach, this data is split into two folds. The first fold, which is the first 10,000 records, is used for learning and building the model. These 10,000 records are equivalent to historical data. The remaining 50,000 records, which are more recent than the first 10,000, are used for testing. Each column in this testing data set is equivalent to the acquirable context attribute at point of inference.

In deployment, an authentic situation is simulated by incorporating missing data, set at rates of 20%, 40%, 60%, 80% and 100% respectively, into the test data sets as explained in 5.2.1 Simulating a Real Situation.

### 5.3.3 The CATE Framework Approach

For the CATE framework approach, we use multiple inference models. Each inference model is created for each set of influential context attributes; each inference model matches the set of context attributes referenced when missing sensory data is detected. This is different from the single model approach that builds only the one inference model that is used for all cases.

For the implementation of the CATE framework, we define a core module called RoadLink that acts as both a Data Access Object (DAO) and a service provider. As can be seen in Figure 4-13, in the CATE framework approach, each road segment has its own RoadLink module. Each RoadLink module can communicate with other RoadLink module (and especially the RoadLink module of a connected road segment) to send traffic information. This is necessary when the RoadLink module of a connected road segment requests traffic data to use as an input to infer the traffic congestion degree. For example, when the sensory traffic data of an observed road segment is not available, the program starts trying to collect available influential context attributes based on the list of predefined influential context attributes. For our evaluation, the context attributes are *day*, *time*, *rain data* and the *sensory traffic data*

*from connected road*. A matching built inference model is chosen according to the available contexts at the time of calculation (inference).

Figure 5-14 helps to explain the implementation of the CATE framework. It shows the part of our Groovy programming code [127] that acts as the main method that calls other sub-methods (this is to simplify the code and make it more readable). The code shown in Figure 5-14 is the core piece of code; the sub-methods can be found in an Appendix B.

```
def process() {

    init()    // learn and build inference models from historical data
              // then save the built models for inferring later


    // create object for raw testing data
    def dataFile = new File("./ex_data/${name}sensor.and.context.csv")

    // read data record by record (equivalent to real situation that the data come every 5 minutes.)
    dataFile.eachLine {line, number ->

        if(number == 1) return

        getData(line, number)     // extract acquirable context to context attribute(s)
        randomMissing(missRate)   // randomize missing sensory traffic data of observed road and
                                  // other context attributes according to specified missing rate
        infer()   // if  the sensory traffic data of observed road is missing
                  // select suitable model based on available context attribute(s)
                  // load the matching model and infer traffic congestion degree
        reportTraffic()   // report actual (sensory traffic data is not missing) or inferred traffic congestion
degree
        saveHistory()   // save new incoming actual data for relearning in the future
        if(NeedToReLearn) learn(algorithm, numberOfRecordToLearn)   // check if it reach the time to
relearn
    }  // loop back to read next record
    summarize()  // generate summary report for experiment and evaluation purpose

}
```

**Figure 5-14: Extract from the Groovy programming code for the CATE framework implementation**

The implementation of the CATE framework approach starts with the learning process to derive the inference model(s) for the various context attribute sets. We use historical data to build the models. We choose the J48 machine learning algorithm (which we evaluated in Chapter 4) to build the inference models as it is fast, gives high accuracy and is suitable for our data. After learning the historical data, different inference models are obtained for each context attribute set. The number of all

possible inference models (IM) is equal to $\sum_{k=1}^{n} {}^{n}C_{k}$ if we have *n* influential context attributes in total and *k* is the number of member(s) in each context attribute set. Based on the information presented in Table 5-1, where the context attributes are *school break, rain, day, time* and *the traffic of connected road*, in Table 5-3 we present examples of context attribute sets for inferring the traffic congestion degree of road segment 2718 when applying five context attributes (if applying six context attributes, the number of context attribute set is more).

**Table 5-3: Context attribute sets for inferring the traffic of road segment ID 2718 (when using five context attributes)**

| Context attribute set i | Member(s) in context attribute set | Inference Model (IM) *m* |
|---|---|---|
| CtxSet1 | School | $IM_1$ |
| CtxSet2 | Rain | $IM_2$ |
| CtxSet3 | Day | $IM_3$ |
| CtxSet4 | 1401 | $IM_4$ |
| CtxSet5 | Time | $IM_5$ |
| CtxSet6 | School, rain | $IM_6$ |
| CtxSet7 | School, day | $IM_7$ |
| CtxSet8 | Rain, day | $IM_8$ |
| CtxSet9 | School, 1401 | $IM_9$ |
| CtxSet10 | Rain, 1401 | $IM_{10}$ |
| CtxSet11 | Day, 1401 | $IM_{11}$ |
| CtxSet12 | Rain, time | $IM_{12}$ |
| CtxSet13 | School, time | $IM_{13}$ |
| CtxSet14 | Time, 1401 | $IM_{14}$ |
| CtxSet15 | Day, time | $IM_{15}$ |
| CtxSet16 | School, rain, day | $IM_{16}$ |
| CtxSet17 | School, rain, 1401 | $IM_{17}$ |
| CtxSet18 | Rain, day, 1401 | $IM_{18}$ |
| CtxSet19 | School, day, 1401 | $IM_{19}$ |
| CtxSet20 | Rain, time, 1401 | $IM_{20}$ |
| CtxSet21 | School, time, rain | $IM_{21}$ |
| CtxSet22 | School, time, 1401 | $IM_{22}$ |
| CtxSet23 | Day, time, 1401 | $IM_{23}$ |
| CtxSet24 | Day, time, rain | $IM_{24}$ |
| CtxSet25 | Day, time, school | $IM_{25}$ |
| CtxSet26 | School, rain, day, 1401 | $IM_{26}$ |
| CtxSet27 | School, rain, time, 1401 | $IM_{27}$ |
| CtxSet28 | Day, time, 1401, rain | $IM_{28}$ |
| CtxSet29 | Day, time, 1401, school | $IM_{29}$ |
| CtxSet30 | Day, time, rain, school | $IM_{30}$ |
| CtxSet31 | Day, time, 1401, rain, school | $IM_{31}$ |

From the example in Table 5-3, the attribute *1401* represents the traffic congestion degree of road segment 1401 that is connected to road segment 2718 (the segment that

is to be inferred). (Table 5-1 presents an explanation of other terms). These context attribute sets are used as the input for the machine learning module to create a specific inference model for each specific context attribute set. For the selected context attributes in Table 5-3, we obtain 31 inference models. If there were more context attributes applied, more inference models would be created.

After building a variety of inference models, we evaluate the models with stratified 10-fold cross validation [134], similar to the single model approach. We build and evaluate the models ten times and then calculate the average and standard deviation of the time required to build the models. Repeating the task for ten times and calculating an average is because the data for testing differs in each time.

From Figure 5-14 it can be seen that after the inference models are built, these models are saved for later use in inference phase. The inference models will be rebuilt when the pre-defined relearning time is reached. During deployment, when the framework detects the absence of sensory traffic data, the system determines the available context attributes and then uses this information to determine which of the saved inference models to use. The inference model that matches the list of current available influential context attribute(s) is selected. The chosen inference model is then used to infer the traffic congestion degree to compensate for the missing sensory traffic data of the observed road segment.

In our evaluation of the CATE framework, like the mode and single model approaches, we use a total of 60,000 records. This total data is divided into historical data and testing data. The first fold (the first 10,000 records) is used for learning and building a model. These 10,000 records represent historical data. The remaining 50,000 records, which are more recent than the first 10,000, are used for testing and equate to the acquirable context attributes available at the point of inference.

Similar to the mode approach and single model approach, to evaluate the CATE framework, each column of the test data set is simulated to be missing data at rates of 20%, 40%, 60%, 80% and 100% respectively, as explained in 5.2.1 Simulating a Real Situation.

Because the CATE framework approach incorporates a relearning function at pre-defined time periods, in this experiment we set our program to relearn and rebuild the

models after every 10,000 records. In practice, this number is equivalent to relearning and rebuilding models on a monthly basis. Thus the 50,000 records for testing are separated into five folds, 10,000 records per fold, and offer five opportunities for relearning. The optimal time period for relearning is a subject left for future research.

The system keeps new incoming actual data to facilitate relearning in the future. Any records that contain empty fields (such as missing sensory data) are discarded. Over time, as new data arrives, the historical data used for relearning and building new models will grow. The system will improve itself automatically as more actual data is collected. A description of the relearning process and its justification is found in 4.4.3 in Chapter 4.

## *5.4 Evaluation Results*

After running the evaluation using the three approaches, in this section we report the results of each experiment and the comparison between the three approaches. For the mode approach, we focus on accuracy as the main measure as the mode approach does not involve generating an inference model. Nonetheless, we do note the time required to generate the substitute data. For the single model and the CATE framework approaches, both the accuracy and the time taken to learn and generate the inference models are reported. Accuracy is defined as the percentage of instances classified correctly, which is calculated by the following equation:

$$Accuracy = \left(\frac{total\ number\ of\ correctly\ infered\ traffic\ congestion\ degree}{total\ number\ of\ traffic\ congestion\ degree\ needed\ to\ be\ inferred}\right) \times 100$$

### 5.4.1 Results from the Mode Approach

Under the mode approach, any missing sensory traffic data of a road segment is replaced by the mode value. The mode is calculated from historical data (the first 10,000 records of our total data as explained in Section 4.3.1 and Section 0). The remaining 50,000 records are treated as test data and incorporate random missing data

at rates of 20%, 40%, 60%, 80% and 100%. Whenever the sensory data of a relevant road segment is not available, the mode is then reported as the inferred traffic congestion degree.

In this section, we start by showing the data distribution of the traffic congestion degree for each relevant road segment to explain the data distribution and to show the mode for each road segment. The tables and graphs demonstrating the accuracy of using the mode approach to create the substitute traffic data are then presented.

Road segment 1206 data distribution

**Table 5-4: Road segment 1206 data distribution and mode**

| | L | M | H | Total | % Mode (L) |
|---|---|---|---|---|---|
| 1-10000 (Historical data) | 8900 | 936 | 164 | 10000 | 89.00% |
| 10001-60000 (Test data) | 43432 | 6012 | 556 | 50000 | 86.86% |

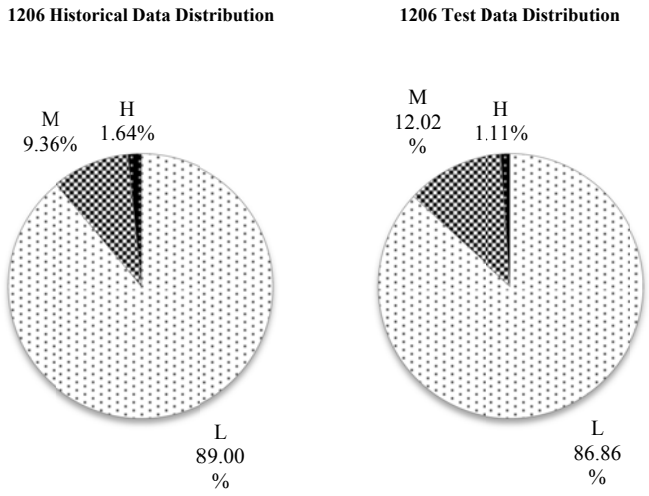



**Figure 5-15: Road segment 1206 data distribution**

From the historical data, the mode or the traffic congestion degree that has the highest frequency for road segment 1206 is *L*.

Road segment 2613 data distribution

**Table 5-5: Road segment 2613 data distribution and mode**

| | L | M | H | Total | % Mode (L) |
|---|---|---|---|---|---|
| 1-10000 (Historical data) | 7189 | 2433 | 378 | 10000 | 71.89% |
| 10001-60000 (Test data) | 37712 | 10281 | 2007 | 50000 | 75.42% |

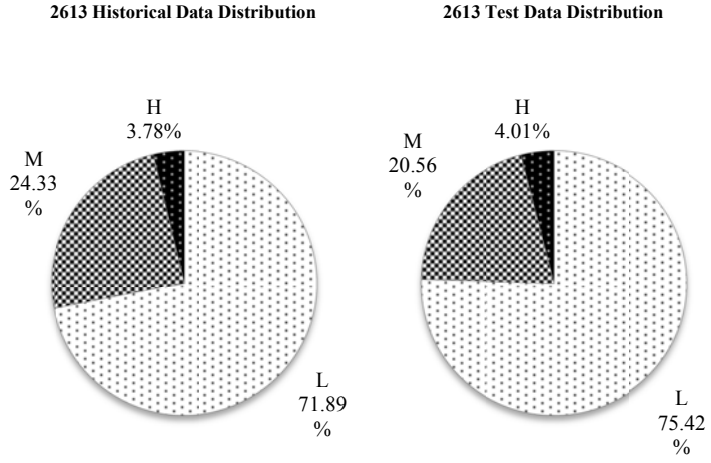



**Figure 5-16: Road segment 2613 data distribution**

From the historical data, the mode or the traffic congestion degree that has the highest frequency for road segment 2613 is *L*.

Road segment 2718 data distribution

**Table 5-6: Road segment 2718 data distribution and mode**

| | L | M | H | Total | % Mode (L) |
|---|---|---|---|---|---|
| 1-10000 (Historical data) | 8244 | 1740 | 16 | 10000 | 82.44% |
| 10001-60000 (Test data) | 43417 | 6562 | 21 | 50000 | 86.83% |

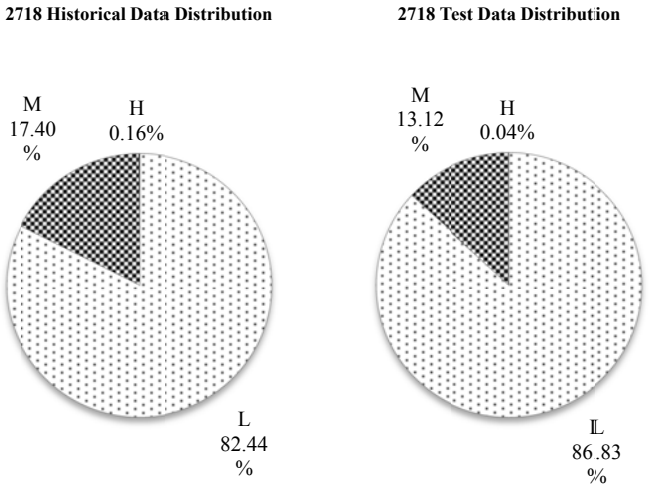



**Figure 5-17: Road segment 2718 data distribution**

From the historical data, the mode or the traffic congestion degree that has the highest frequency for road segment 2718 is *L*.

The following table presents the degree of accuracy obtained when using the simplest static method (that is, replacing the missing value with the mode) to infer missing sensory traffic data. The mean and standard deviation at each level of missing data for each road link are presented.

**Table 5-7: Accuracy of the mode approach at specific missing data rates**

| Road Link | Accuracy (%) at specific missing data rates (%) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Miss 20% | | Miss 40% | | Miss 60% | | Miss 80% | | Miss 100% | |
| | Mean | S.D. | Mean | S.D. | Mean | S.D. | Mean | S.D. | Mean | S.D. |
| 1206 | 86.83 | 0.23 | 86.84 | 0.17 | 86.86 | 0.14 | 86.82 | 0.08 | 86.86 | 0.00 |
| 2613 | 75.41 | 0.35 | 75.33 | 0.20 | 75.39 | 0.13 | 75.41 | 0.10 | 75.42 | 0.00 |
| 2718 | 86.94 | 0.37 | 86.94 | 0.16 | 86.81 | 0.14 | 86.82 | 0.06 | 86.83 | 0.00 |

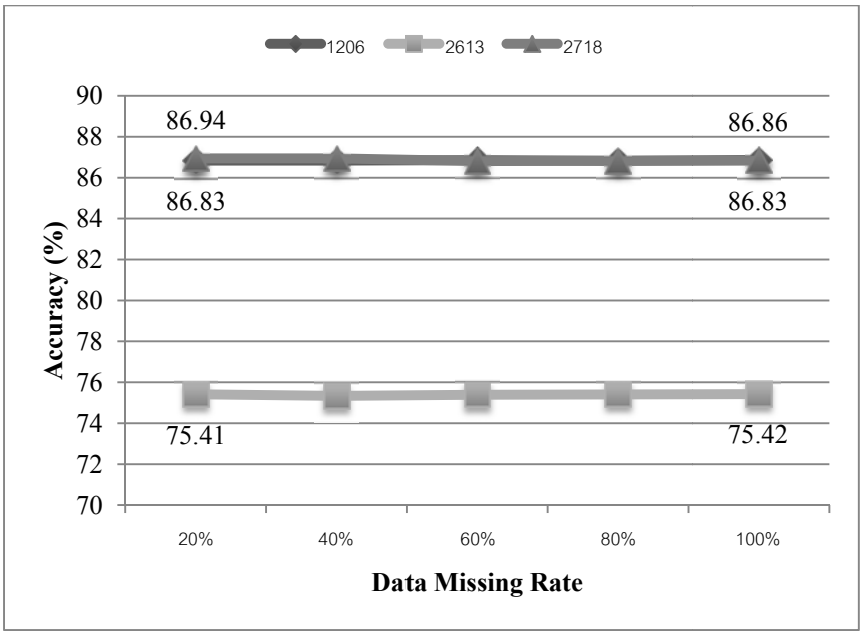



**Figure 5-18: Accuracy of mode approach at specific missing data rates**

The difference between active and non-active periods based on the level of traffic congestion at particular times during each 24 hours was described in Section 5.1.2. In Table 5-8 and Table 5-9 we present results based on both active and non-active time periods.

**Table 5-8: Accuracy of the mode approach at specific missing data rates (*active period*)**

| Road Link | Active period accuracy (%) at specific missing data rates (%) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Miss 20% | | Miss 40% | | Miss 60% | | Miss 80% | | Miss 100% | |
| | Mean | S.D. | Mean | S.D. | Mean | S.D. | Mean | S.D. | Mean | S.D. |
| 1206 | 77.24 | 0.31 | 77.21 | 0.32 | 77.24 | 0.24 | 77.20 | 0.17 | 77.22 | 0.00 |
| 2613 | 51.18 | 0.73 | 51.13 | 0.34 | 51.19 | 0.27 | 51.28 | 0.17 | 51.26 | 0.00 |
| 2718 | 74.80 | 0.63 | 74.97 | 0.35 | 74.75 | 0.24 | 74.72 | 0.11 | 74.75 | 0.00 |

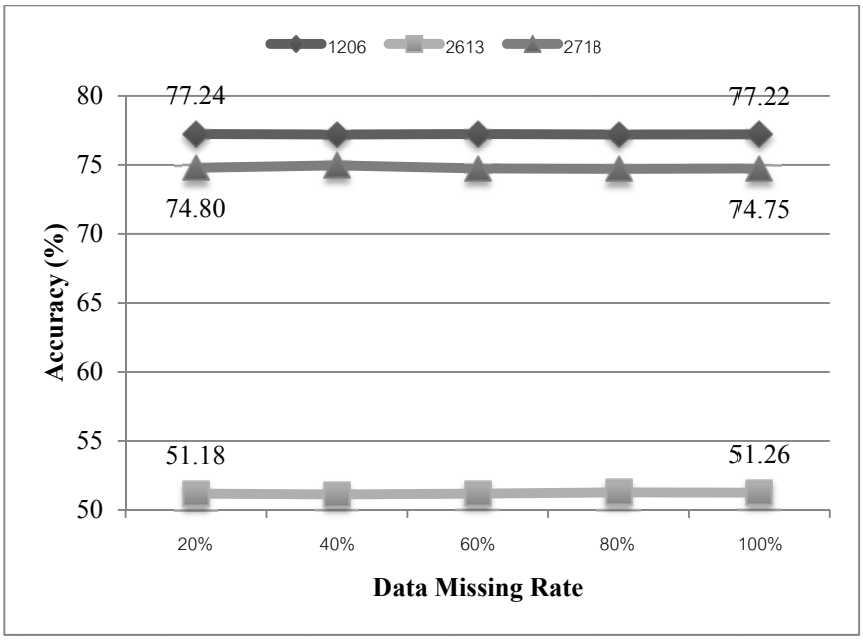



**Figure 5-19: Accuracy of the mode approach at specific missing data rates (*active period*)**

**Table 5-9: Accuracy of the mode approach at specific missing data rates (*non-active period*)**

| Road Link | Non-active period accuracy (%) at specific missing data rates (%) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Miss 20% | | Miss 40% | | Miss 60% | | Miss 80% | | Miss 100% | |
| | Mean | S.D. | Mean | S.D. | Mean | S.D. | Mean | S.D. | Mean | S.D. |
| 1206 | 96.36 | 0.32 | 96.26 | 0.10 | 96.33 | 0.05 | 96.31 | 0.05 | 96.35 | 0.00 |
| 2613 | 99.14 | 0.10 | 99.15 | 0.07 | 99.21 | 0.05 | 99.19 | 0.01 | 99.20 | 0.00 |
| 2718 | 98.76 | 0.14 | 98.74 | 0.10 | 98.71 | 0.07 | 98.73 | 0.03 | 98.73 | 0.00 |

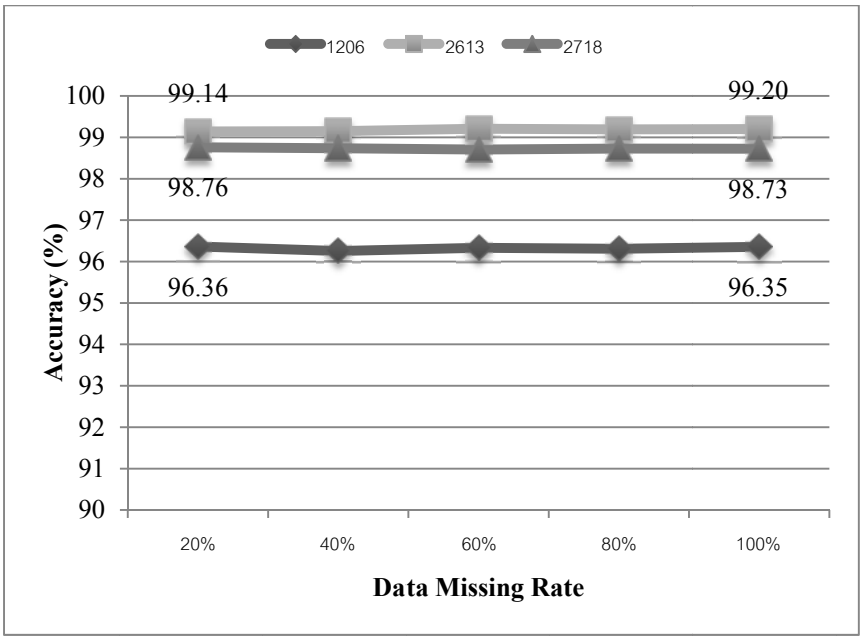



**Figure 5-20: Accuracy of the mode approach at specific missing data rates (*non-active period*)**

Discussion

It is noticeable that the characteristics of the data for these three road segments are similar. Moreover, it can be seen that the accuracy rates remain the same for all the missing data rates because the inferred congestion degree is always the same (that is, the mode value, *L*). The accuracy is almost the same as the value representing the actual traffic congestion degree (*L* in the test data, as can be seen in Figure 5-18)

because when the sensory traffic data is missing, the system will use the mode value to report the inferred traffic congestion degree.

A different picture arises when we consider only the active period (shown in Figure 5-19). During the active period, the accuracy drops considerably compared to the accuracy of all periods (shown in Figure 5-18). This is especially true for road segment 2613, where the accuracy drops to 50%. This drop in accuracy arises because the *M* and *H* traffic congestion degrees occur more frequently in the active period than in the non-active period. The number of *M* and *H* situations are not correctly inferred because the system always replaces the missing traffic data with the mode value (in this case, *L*). Using mode approach is not appropriate even though the infer time is less than 0.2 μs.. Considering when the actual traffic data is "H" but the system infers the missing traffic data as *L*. This drop in accuracy has ramifications. In practical terms, the difference between an *H* traffic congestion degree and an *L* traffic congestion degree is substantial. If a traffic system infers the traffic congestion degree as *L* when the actual traffic congestion degree is *H*, it would report erroneous information to drivers and could cause poor route choices.

### 5.4.2 Results from the Single Model Approach

In this section, using the single model approach, we evaluate the time taken to build a model and the accuracy of the model's outputs. We build and evaluate the model ten times and calculate the average and standard deviation of both the build time and the accuracy of the results. As we explained earlier, in the single model approach only one inference model is built for all cases. The missing context attribute(s) in the test data (which is equivalent to the context attribute at the current time or point of inference) is manipulated by the DBI method (explained in Section 5.3.2). Similar to the mode approach in Section 5.4.1, the non-active period is anytime on Saturday and Sunday and from 2300 hours to 0600 hours Monday through to Friday. The remaining time is the active period in which traffic conditions fluctuate (see the distribution of traffic data in Section 5.1.2.)

For the model building in an experiment of single model approach, we build and evaluate the model ten times and calculate the average and standard deviation (S.D.) of accuracy and model building time. The model evaluation is to test the accuracy of

the obtained model before deployment. The results for each road segment are presented in Table 5-10.

**Table 5-10: Accuracy and model building time of the single model approach**

| Road Link | Accuracy (%) | | Model Building Time (μs) | |
|---|---|---|---|---|
| | Average | S.D. | Average | S.D. |
| 1206 | 93.78 | 0.08 | 8318 | 808 |
| 2613 | 87.17 | 0.12 | 10280 | 1853 |
| 2718 | 98.99 | 0.01 | 6980 | 1297 |

Once the inference model is built, it is used to infer the missing sensory traffic data of the observed road segments. A real situation is simulated, as described in Section 5.2.1, by using missing data. In this section, we present the results of the deployment when the missing data rates are set to 20%, 40%, 60%, 80% and 100% respectively. We ran the experiment ten times and show the average of those ten runs in this section. Table 5-11 and the graph in Figure 5-21 show the average time used to create the inferred traffic congestion degree when the sensory traffic data of a particular road segment is missing. Table 5-12 shows the average accuracy and standard deviation of the results at different missing data rates for each road segment. We also show the results for the active and non-active periods in Table 5-13 and Table 5-14.

**Table 5-11: Inferring times of the single model approach at specific missing data rates**

| Road Segment | Inferring time (μs) at specific missing data rates (%) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Miss 20% | | Miss 40% | | Miss 60% | | Miss 80% | | Miss 100% | |
| | Average | S.D. | Average | S.D. | Average | S.D. | Average | S.D. | Average | S.D. |
| 1206 | 1470 | 31 | 1464 | 18 | 1472 | 18 | 1472 | 19 | 1459 | 21 |
| 2613 | 2195 | 44 | 2198 | 27 | 2202 | 26 | 2211 | 22 | 2206 | 34 |
| 2718 | 1092 | 22 | 1092 | 13 | 1097 | 9 | 1090 | 8 | 1088 | 20 |

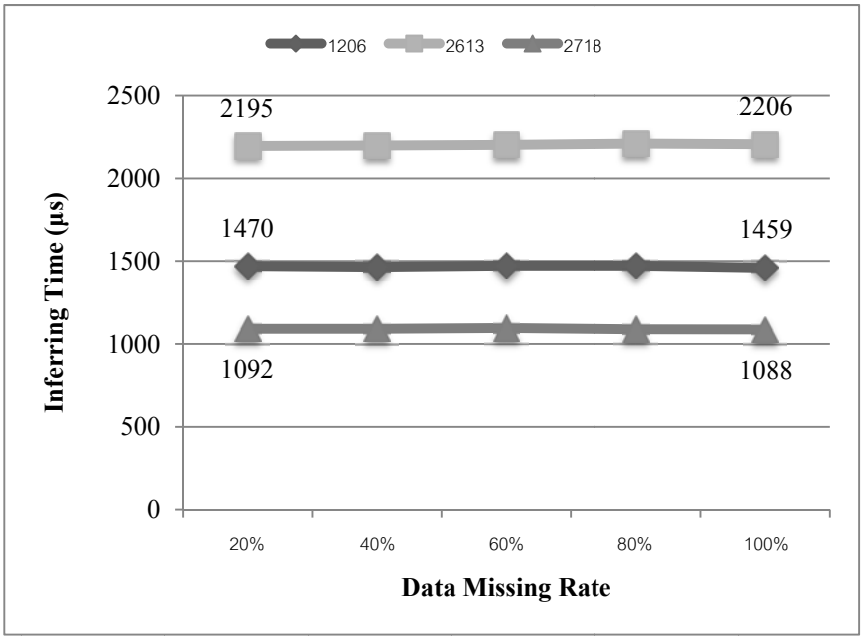



**Figure 5-21: Inference times of the single model approach at specific missing data rates**

**Table 5-12: Accuracy of the single model approach at specific missing data rates**

| Road Link | Accuracy (%) at specific missing data rates (%) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Miss 20% | | Miss 40% | | Miss 60% | | Miss 80% | | Miss 100% | |
| | Average | S.D. | Average | S.D. | Average | S.D. | Average | S.D. | Average | S.D. |
| 1206 | 87.78 | 0.16 | 87.96 | 0.13 | 87.96 | 0.17 | 88.08 | 0.10 | 88.11 | 0.00 |
| 2613 | 79.55 | 0.46 | 79.18 | 0.27 | 78.24 | 0.14 | 77.04 | 0.11 | 75.32 | 0.00 |
| 2718 | 96.48 | 0.08 | 94.15 | 0.24 | 91.66 | 0.11 | 89.34 | 0.10 | 86.83 | 0.00 |

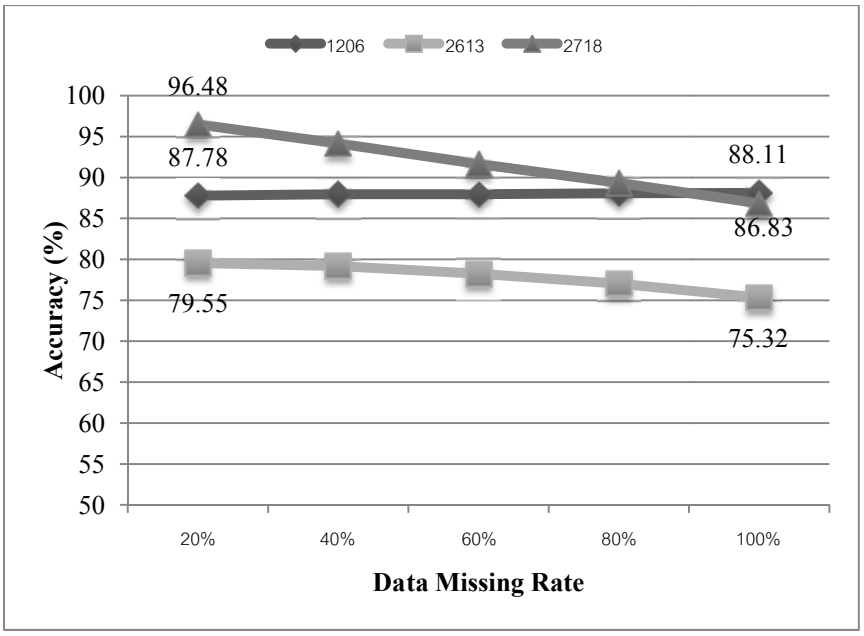



**Figure 5-22: Accuracy of the single model approach at specific missing data rates**

Noticeable in the above graph is the reduction in accuracy at the higher missing data rates for road segments 2613 and 2718. In contrast, road segment 1206 retains a reasonably constant accuracy rate. This may be due to the context attributes, apart from day and time, not having much impact on the traffic conditions of road segment 1206. The missing data rate of 100% means that only day and time context attributes will be used to infer the traffic congestion degrees of these particular road segments.

**Table 5-13: Accuracy of the single model approach at specific missing data rates (*active period*)**

| Road Link | Active period accuracy (%) at specific data missing rate (%) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Miss 20% | | Miss 40% | | Miss 60% | | Miss 80% | | Miss 100% | |
| | Average | S.D. | Average | S.D. | Average | S.D. | Average | S.D. | Average | S.D. |
| 1206 | 79.65 | 0.38 | 79.88 | 0.21 | 79.79 | 0.33 | 79.98 | 0.20 | 79.97 | 0.00 |
| 2613 | 61.43 | 0.96 | 60.17 | 0.55 | 57.93 | 0.27 | 54.98 | 0.17 | 51.04 | 0.00 |
| 2718 | 93.30 | 0.18 | 88.82 | 0.40 | 84.00 | 0.21 | 79.57 | 0.18 | 74.75 | 0.00 |

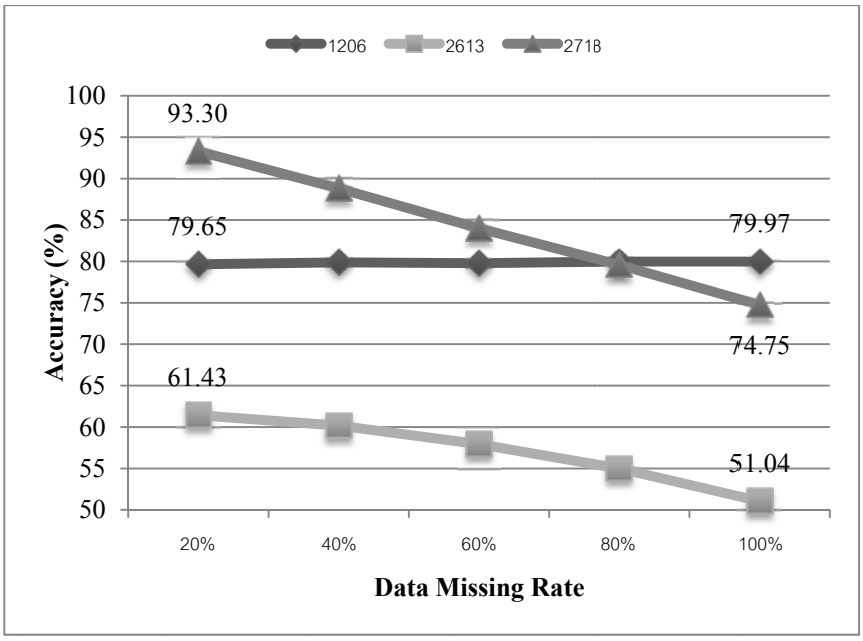



**Figure 5-23: Accuracy of the single model approach at specific missing data rates (*active period*)**

When considering only the active period in the above graph, the accuracy reduction of accuracy is more than the overall accuracy shown in Figure 5-22 when missing data rates are higher.

**Table 5-14: Accuracy of the single model approach at specific missing data rates (*non-active period*)**

| Road Link | Non-Active period accuracy (%) at specific missing data rates (%) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Miss 20% | | Miss 40% | | Miss 60% | | Miss 80% | | Miss 100% | |
| | Average | S.D. | Average | S.D. | Average | S.D. | Average | S.D. | Average | S.D. |
| 1206 | 95.79 | 0.25 | 95.94 | 0.16 | 96.00 | 0.08 | 96.07 | 0.10 | 96.11 | 0.00 |
| 2613 | 97.46 | 0.13 | 97.90 | 0.08 | 98.26 | 0.08 | 98.71 | 0.06 | 99.20 | 0.00 |
| 2718 | 99.62 | 0.08 | 99.41 | 0.09 | 99.17 | 0.06 | 98.96 | 0.03 | 98.73 | 0.00 |

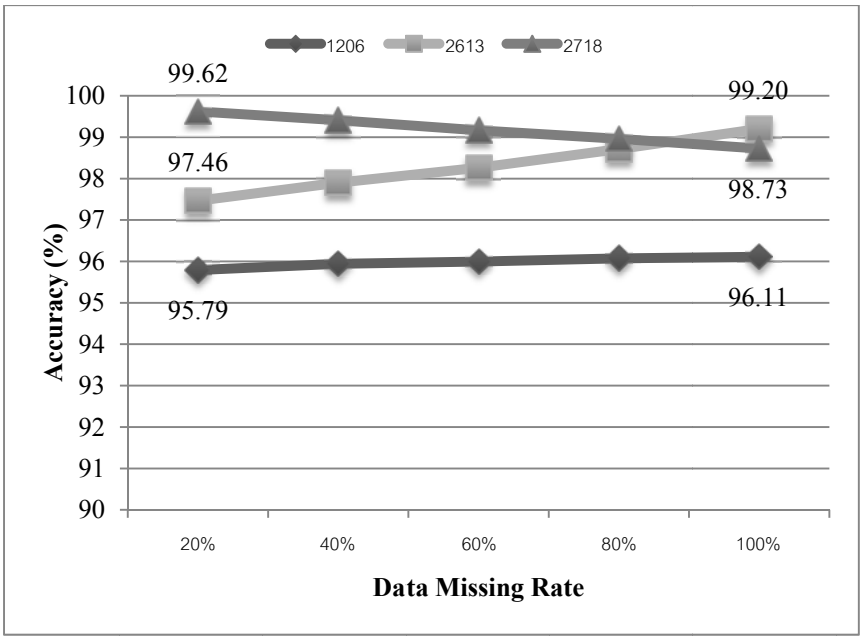



**Figure 5-24: Accuracy of the single model approach at specific missing data rates (*non-active period*)**

In Figure 5-24, the level of accuracy for the non-active period, when using the single model approach, varies for all three road segments. The accuracy for road segment 2613 increases at a higher missing data rate. This may be caused by the DBI method that the J48 algorithm uses to manipulate the missing context attributes [103]. DBI relies on the frequency-based probability of the missing attribute to replace the missing value, which, in the case of the non-active period, usually leads to the result *L*. Because most of the actual traffic congestion degrees in the non-active period are *L*, DBI thus helps to increase the estimation accuracy; in this case, at the higher rates of

missing data. However, the increase in accuracy is only slight, with no practical ramification.

### 5.4.3 Results from the CATE Framework Approach

The process started by building multiple models as explained in Section 4.4.1. The built models are evaluated to test the accuracy. To evaluate the models for our proposed CATE framework, we also use the stratified 10-fold cross validation method for evaluation. We build and evaluate the models ten times and then calculate the average and standard deviation for both accuracy and the average model building time. For the CATE framework, several inference models are built as explained in Section 4.4.1 and in Section 4.4.3 in Chapter 4.

Table 5-16, Table 5-15 and Table 5-17 show the results of the framework's accuracy evaluation using average and standard deviation. We also include the average model building time and standard deviation of the different context attribute combinations using J48 machine learning algorithms for the observed road segments with road segment IDs 2613, 1206 and 2718 respectively. To make the results more understandable and to facilitate comparison, we place the graphs of accuracy and model building time inside the tables.

**Table 5-15: Averages of accuracy and model building times generated from different inference models for road segment 1206**

| CtxSet | Context Attributes Applied | Accuracy (%) | | Models Building Time (µs) | |
|---|---|---|---|---|---|
| | | Mean | S.D. | Mean | S.D. |
| 1 | {Day} | 89.00 | 0.00 | 3201 | 493 |
| 2 | {Time} | 89.71 | 0.10 | 3101 | 251 |
| 3 | {Rain} | 89.00 | 0.00 | 3378 | 588 |
| 4 | {School} | 89.00 | 0.00 | 2908 | 418 |
| 5 | {1211} | 89.29 | 0.00 | 3052 | 164 |
| 6 | {1403} | 89.00 | 0.00 | 3413 | 478 |
| 7 | {Day,Time} | 90.61 | 0.11 | 4147 | 338 |
| 8 | {Day,Rain} | 89.00 | 0.00 | 4677 | 600 |
| 9 | {Day,School} | 89.00 | 0.00 | 4212 | 443 |
| 10 | {Day,1211} | 89.46 | 0.01 | 5672 | 1970 |
| 11 | {Day,1403} | 89.00 | 0.00 | 4969 | 570 |
| 12 | {Time,Rain} | 89.83 | 0.10 | 4174 | 441 |
| 13 | {Time,School} | 89.71 | 0.10 | 4122 | 611 |
| 14 | {Time,1211} | 92.01 | 0.08 | 4553 | 740 |
| 15 | {Time,1403} | 90.89 | 0.05 | 4627 | 492 |
| 16 | {Rain,School} | 89.00 | 0.00 | 4409 | 473 |
| 17 | {Rain,1211} | 89.35 | 0.04 | 5574 | 1782 |
| 18 | {Rain,1403} | 89.00 | 0.00 | 4576 | 452 |
| 19 | {School,1211} | 89.29 | 0.00 | 4215 | 432 |
| 20 | {School,1403} | 89.00 | 0.00 | 4110 | 669 |
| 21 | {1211,1403} | 89.25 | 0.02 | 5113 | 647 |
| 22 | {Day,Time,Rain} | 90.48 | 0.13 | 5344 | 2196 |
| 23 | {Day,Time,School} | 90.61 | 0.11 | 4151 | 776 |
| 24 | {Day,Time,1211} | 92.70 | 0.06 | 4917 | 1192 |
| 25 | {Day,Time,1403} | 92.08 | 0.12 | 4483 | 851 |
| 26 | {Day,Rain,School} | 89.00 | 0.00 | 4256 | 431 |
| 27 | {Day,Rain,1211} | 89.46 | 0.01 | 4983 | 326 |
| 28 | {Day,Rain,1403} | 89.00 | 0.00 | 4853 | 861 |
| 29 | {Day,School,1211} | 89.46 | 0.01 | 4689 | 456 |
| 30 | {Day,School,1403} | 89.00 | 0.00 | 4458 | 495 |
| 31 | {Day,1211,1403} | 89.58 | 0.02 | 5370 | 815 |
| 32 | {Time,Rain,School} | 89.83 | 0.10 | 3787 | 348 |
| 33 | {Time,Rain,1211} | 92.11 | 0.07 | 4515 | 482 |
| 34 | {Time,Rain,1403} | 91.07 | 0.06 | 3987 | 464 |
| 35 | {Time,School,1211} | 92.01 | 0.08 | 4204 | 665 |
| 36 | {Time,School,1403} | 90.89 | 0.05 | 4164 | 604 |
| 37 | {Time,1211,1403} | 92.29 | 0.12 | 4459 | 414 |
| 38 | {Rain,School,1211} | 89.35 | 0.04 | 4934 | 955 |
| 39 | {Rain,School,1403} | 89.00 | 0.00 | 4551 | 723 |
| 40 | {Rain,1211,1403} | 89.33 | 0.02 | 5303 | 653 |
| 41 | {School,1211,1403} | 89.25 | 0.02 | 4427 | 527 |
| 42 | {Day,Time,Rain,School} | 90.48 | 0.13 | 4635 | 965 |
| 43 | {Day,Time,Rain,1211} | 92.83 | 0.06 | 5514 | 575 |
| 44 | {Day,Time,Rain,1403} | 92.11 | 0.11 | 4922 | 576 |
| 45 | {Day,Time,School,1211} | 92.70 | 0.06 | 4833 | 625 |
| 46 | {Day,Time,School,1403} | 92.08 | 0.12 | 4692 | 342 |
| 47 | {Day,Time,1211,1403} | 93.75 | 0.08 | 6136 | 1866 |
| 48 | {Day,Rain,School,1211} | 89.46 | 0.01 | 5425 | 511 |
| 49 | {Day,Rain,School,1403} | 89.00 | 0.00 | 5315 | 505 |
| 50 | {Day,Rain,1211,1403} | 89.65 | 0.01 | 5787 | 617 |
| 51 | {Day,School,1211,1403} | 89.58 | 0.02 | 5567 | 611 |
| 52 | {Time,Rain,School,1211} | 92.11 | 0.07 | 4789 | 477 |
| 53 | {Time,Rain,School,1403} | 91.07 | 0.06 | 4831 | 738 |
| 54 | {Time,Rain,1211,1403} | 92.39 | 0.09 | 5342 | 768 |
| 55 | {Time,School,1211,1403} | 92.29 | 0.12 | 4965 | 570 |
| 56 | {Rain,School,1211,1403} | 89.33 | 0.02 | 5757 | 564 |
| 57 | {Day,Time,Rain,School,1211} | 92.83 | 0.06 | 6098 | 376 |
| 58 | {Day,Time,Rain,School,1403} | 92.11 | 0.11 | 5472 | 2130 |
| 59 | {Day,Time,Rain,1211,1403} | 93.78 | 0.08 | 6742 | 977 |
| 60 | {Day,Time,School,1211,1403} | 93.75 | 0.08 | 6065 | 1011 |
| 61 | {Day,Rain,School,1211,1403} | 89.65 | 0.01 | 7978 | 905 |
| 62 | {Time,Rain,School,1211,1403} | 92.39 | 0.09 | 6553 | 434 |
| 63 | {Day,Time,Rain,School,1211,1403} | 93.78 | 0.08 | 7222 | 802 |
| | | **Total Model Building Time (µs)** | | **304678** | |

**Table 5-16: Averages of accuracy and model building times generated from different inference models for road segment 2613**

| CtxSet | Context Attributes Applied | Accuracy (%) | | Models Building Time (µs) | |
|---|---|---|---|---|---|
| | | Mean | S.D. | Mean | S.D. |
| 1 | {Day} | 71.89 | 0.00 | 3236 | 605 |
| 2 | {Time} | 74.60 | 0.07 | 3281 | 275 |
| 3 | {Rain} | 71.93 | 0.01 | 3215 | 366 |
| 4 | {School} | 71.89 | 0.00 | 2795 | 305 |
| 5 | {3015} | 78.52 | 0.00 | 3174 | 414 |
| 6 | {2614} | 78.95 | 0.00 | 3660 | 999 |
| 7 | {Day,Time} | 82.49 | 0.10 | 4613 | 559 |
| 8 | {Day,Rain} | 71.88 | 0.04 | 4205 | 399 |
| 9 | {Day,School} | 71.89 | 0.00 | 4024 | 536 |
| 10 | {Day,3015} | 81.10 | 0.00 | 4558 | 596 |
| 11 | {Day,2614} | 79.26 | 0.01 | 5158 | 671 |
| 12 | {Time,Rain} | 74.75 | 0.10 | 4817 | 1245 |
| 13 | {Time,School} | 74.60 | 0.07 | 4092 | 246 |
| 14 | {Time,3015} | 80.37 | 0.07 | 4915 | 900 |
| 15 | {Time,2614} | 79.72 | 0.09 | 4483 | 390 |
| 16 | {Rain,School} | 71.93 | 0.01 | 4333 | 505 |
| 17 | {Rain,3015} | 78.36 | 0.05 | 5056 | 1721 |
| 18 | {Rain,2614} | 78.90 | 0.02 | 5031 | 559 |
| 19 | {School,3015} | 78.52 | 0.00 | 4579 | 880 |
| 20 | {School,2614} | 78.95 | 0.00 | 4407 | 902 |
| 21 | {3015,2614} | 79.50 | 0.19 | 4657 | 384 |
| 22 | {Day,Time,Rain} | 82.42 | 0.12 | 4476 | 593 |
| 23 | {Day,Time,School} | 82.49 | 0.10 | 4527 | 849 |
| 24 | {Day,Time,3015} | 85.06 | 0.13 | 5083 | 467 |
| 25 | {Day,Time,2614} | 84.47 | 0.14 | 5351 | 887 |
| 26 | {Day,Rain,School} | 71.88 | 0.04 | 4262 | 493 |
| 27 | {Day,Rain,3015} | 81.23 | 0.03 | 4467 | 550 |
| 28 | {Day,Rain,2614} | 79.46 | 0.01 | 4985 | 615 |
| 29 | {Day,School,3015} | 81.10 | 0.00 | 4700 | 973 |
| 30 | {Day,School,2614} | 79.26 | 0.01 | 4890 | 1736 |
| 31 | {Day,3015,2614} | 81.81 | 0.10 | 4784 | 401 |
| 32 | {Time,Rain,School} | 74.75 | 0.10 | 4167 | 747 |
| 33 | {Time,Rain,3015} | 80.48 | 0.09 | 5327 | 1025 |
| 34 | {Time,Rain,2614} | 80.09 | 0.07 | 4844 | 395 |
| 35 | {Time,School,3015} | 80.37 | 0.07 | 4148 | 638 |
| 36 | {Time,School,2614} | 79.72 | 0.09 | 4208 | 506 |
| 37 | {Time,3015,2614} | 82.79 | 0.06 | 4930 | 501 |
| 38 | {Rain,School,3015} | 78.36 | 0.05 | 4105 | 277 |
| 39 | {Rain,School,2614} | 78.90 | 0.02 | 4308 | 288 |
| 40 | {Rain,3015,2614} | 79.57 | 0.17 | 5780 | 1876 |
| 41 | {School,3015,2614} | 79.50 | 0.19 | 4279 | 355 |
| 42 | {Day,Time,Rain,School} | 82.42 | 0.12 | 4549 | 380 |
| 43 | {Day,Time,Rain,3015} | 85.31 | 0.12 | 5860 | 667 |
| 44 | {Day,Time,Rain,2614} | 84.61 | 0.13 | 6034 | 423 |
| 45 | {Day,Time,School,3015} | 85.06 | 0.13 | 5502 | 326 |
| 46 | {Day,Time,School,2614} | 84.47 | 0.14 | 5883 | 389 |
| 47 | {Day,Time,3015,2614} | 87.23 | 0.13 | 6877 | 357 |
| 48 | {Day,Rain,School,3015} | 81.23 | 0.03 | 5186 | 609 |
| 49 | {Day,Rain,School,2614} | 79.46 | 0.01 | 5233 | 690 |
| 50 | {Day,Rain,3015,2614} | 82.34 | 0.07 | 5630 | 440 |
| 51 | {Day,School,3015,2614} | 81.81 | 0.10 | 5923 | 2155 |
| 52 | {Time,Rain,School,3015} | 80.48 | 0.09 | 5139 | 517 |
| 53 | {Time,Rain,School,2614} | 80.09 | 0.07 | 5448 | 638 |
| 54 | {Time,Rain,3015,2614} | 83.22 | 0.08 | 6194 | 715 |
| 55 | {Time,School,3015,2614} | 82.79 | 0.06 | 5682 | 818 |
| 56 | {Rain,School,3015,2614} | 79.57 | 0.17 | 5600 | 662 |
| 57 | {Day,Time,Rain,School,3015} | 85.31 | 0.12 | 6222 | 684 |
| 58 | {Day,Time,Rain,School,2614} | 84.61 | 0.13 | 6657 | 353 |
| 59 | {Day,Time,Rain,3015,2614} | 87.17 | 0.12 | 7623 | 525 |
| 60 | {Day,Time,School,3015,2614} | 87.23 | 0.13 | 7397 | 1163 |
| 61 | {Day,Rain,School,3015,2614} | 82.34 | 0.07 | 7199 | 1384 |
| 62 | {Time,Rain,School,3015,2614} | 83.22 | 0.08 | 7784 | 980 |
| 63 | {Day,Time,Rain,School,3015,2614} | 87.17 | 0.12 | 8601 | 836 |
| | | **Total Model Building Time (µs)** | | **318136** | |

**Table 5-17: Averages of accuracy and model building times generated from different inference models for road segment 2718**

| CtxSet | Context Attributes Applied | Accuracy (%) | | Models Building Time | |
|---|---|---|---|---|---|
| | | Mean | S.D. | Mean | S.D. |
| 1 | {Day} | 82.44 | 0.00 | 3279 | 816 |
| 2 | {Time} | 86.40 | 0.00 | 3088 | 418 |
| 3 | {Rain} | 82.44 | 0.00 | 3685 | 1064 |
| 4 | {School} | 82.44 | 0.00 | 3042 | 730 |
| 5 | {2515} | 98.89 | 0.00 | 3431 | 540 |
| 6 | {1401} | 83.71 | 0.00 | 3317 | 532 |
| 7 | {Day,Time} | 89.99 | 0.07 | 4453 | 482 |
| 8 | {Day,Rain} | 82.34 | 0.06 | 4548 | 713 |
| 9 | {Day,School} | 82.44 | 0.00 | 4350 | 688 |
| 10 | {Day,2515} | 98.89 | 0.00 | 4557 | 666 |
| 11 | {Day,1401} | 83.71 | 0.00 | 4800 | 589 |
| 12 | {Time,Rain} | 86.50 | 0.02 | 4810 | 836 |
| 13 | {Time,School} | 86.40 | 0.00 | 4191 | 508 |
| 14 | {Time,2515} | 98.89 | 0.00 | 5311 | 1674 |
| 15 | {Time,1401} | 86.55 | 0.07 | 4348 | 819 |
| 16 | {Rain,School} | 82.44 | 0.00 | 4159 | 614 |
| 17 | {Rain,2515} | 98.89 | 0.00 | 5075 | 894 |
| 18 | {Rain,1401} | 83.70 | 0.02 | 5211 | 520 |
| 19 | {School,2515} | 98.89 | 0.00 | 4455 | 586 |
| 20 | {School,1401} | 83.71 | 0.00 | 4415 | 678 |
| 21 | {2515,1401} | 98.89 | 0.00 | 4701 | 790 |
| 22 | {Day,Time,Rain} | 90.11 | 0.05 | 4533 | 864 |
| 23 | {Day,Time,School} | 89.99 | 0.07 | 4926 | 1042 |
| 24 | {Day,Time,2515} | 98.85 | 0.02 | 4747 | 815 |
| 25 | {Day,Time,1401} | 90.22 | 0.09 | 4895 | 1821 |
| 26 | {Day,Rain,School} | 82.34 | 0.06 | 4344 | 457 |
| 27 | {Day,Rain,2515} | 98.89 | 0.00 | 4577 | 442 |
| 28 | {Day,Rain,1401} | 83.68 | 0.04 | 5191 | 962 |
| 29 | {Day,School,2515} | 98.89 | 0.00 | 4421 | 359 |
| 30 | {Day,School,1401} | 83.71 | 0.00 | 4183 | 674 |
| 31 | {Day,2515,1401} | 98.87 | 0.01 | 5515 | 2255 |
| 32 | {Time,Rain,School} | 86.50 | 0.02 | 4120 | 614 |
| 33 | {Time,Rain,2515} | 98.91 | 0.01 | 4668 | 639 |
| 34 | {Time,Rain,1401} | 86.61 | 0.04 | 4245 | 994 |
| 35 | {Time,School,2515} | 98.89 | 0.00 | 4472 | 766 |
| 36 | {Time,School,1401} | 86.55 | 0.07 | 3626 | 275 |
| 37 | {Time,2515,1401} | 98.96 | 0.01 | 4790 | 665 |
| 38 | {Rain,School,2515} | 98.89 | 0.00 | 4470 | 435 |
| 39 | {Rain,School,1401} | 83.70 | 0.02 | 4531 | 456 |
| 40 | {Rain,2515,1401} | 98.89 | 0.00 | 4779 | 762 |
| 41 | {School,2515,1401} | 98.89 | 0.00 | 3764 | 312 |
| 42 | {Day,Time,Rain,School} | 90.11 | 0.05 | 4841 | 733 |
| 43 | {Day,Time,Rain,2515} | 98.87 | 0.02 | 5117 | 662 |
| 44 | {Day,Time,Rain,1401} | 90.60 | 0.09 | 4486 | 372 |
| 45 | {Day,Time,School,2515} | 98.85 | 0.02 | 5347 | 690 |
| 46 | {Day,Time,School,1401} | 90.22 | 0.09 | 4590 | 698 |
| 47 | {Day,Time,2515,1401} | 98.97 | 0.01 | 4895 | 550 |
| 48 | {Day,Rain,School,2515} | 98.89 | 0.00 | 4992 | 441 |
| 49 | {Day,Rain,School,1401} | 83.68 | 0.04 | 6283 | 1579 |
| 50 | {Day,Rain,2515,1401} | 98.87 | 0.01 | 5423 | 506 |
| 51 | {Day,School,2515,1401} | 98.87 | 0.01 | 4904 | 798 |
| 52 | {Time,Rain,School,2515} | 98.91 | 0.01 | 5038 | 522 |
| 53 | {Time,Rain,School,1401} | 86.61 | 0.04 | 5007 | 1457 |
| 54 | {Time,Rain,2515,1401} | 98.99 | 0.02 | 5407 | 1719 |
| 55 | {Time,School,2515,1401} | 98.96 | 0.01 | 5127 | 1533 |
| 56 | {Rain,School,2515,1401} | 98.89 | 0.00 | 5599 | 2024 |
| 57 | {Day,Time,Rain,School,2515} | 98.87 | 0.02 | 6140 | 2138 |
| 58 | {Day,Time,Rain,School,1401} | 90.60 | 0.09 | 5508 | 785 |
| 59 | {Day,Time,Rain,2515,1401} | 98.99 | 0.01 | 5896 | 1447 |
| 60 | {Day,Time,School,2515,1401} | 98.97 | 0.01 | 5217 | 520 |
| 61 | {Day,Rain,School,2515,1401} | 98.87 | 0.01 | 6335 | 675 |
| 62 | {Time,Rain,School,2515,1401} | 98.99 | 0.02 | 6611 | 1740 |
| 63 | {Day,Time,Rain,School,2515,1401} | 98.99 | 0.01 | 6657 | 1462 |
| | | **Total Model Building Time (µs)** | | **299445** | |

After building multiple inference models for every context attribute sets, these inference models are used to infer traffic data for the observed road segments that will compensate for the missing sensory data. A real situation is simulated as described in Section 5.2.1. Similar to the single model approach, the results when the missing data rate is set to 20%, 40%, 60%, 80% and 100% respectively are presented in this section. We also ran the experiment ten times and present the average accuracy and average inferring time of the ten runs in this section. Table 5-18 and the graph in Figure 5-25 illustrate the average time used to infer the traffic congestion degree. Table 5-12 and the graph in Figure 5-26 show the average accuracy and standard deviation (S.D.) at the different missing data rates for each road segment. In addition, similar to the mode and single model approaches, we also present the results for both active and non-active periods respectively in Table 5-20 and Table 5-21 and illustrate them in Figure 5-27 and Figure 5-28.

**Table 5-18: Inferring times of the CATE framework approach at specific missing data rates**

| Road Link | Inference times (μs) at specific missing data rates (%) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Miss 20% | | Miss 40% | | Miss 60% | | Miss 80% | | Miss 100% | |
| | Mean | S.D. | Mean | S.D. | Mean | S.D. | Mean | S.D. | Mean | S.D. |
| 1206 | 1761 | 38 | 1483 | 45 | 1258 | 20 | 1127 | 23 | 1020 | 20 |
| 2613 | 2360 | 100 | 1938 | 75 | 1558 | 49 | 1343 | 119 | 1141 | 16 |
| 2718 | 1346 | 66 | 1296 | 36 | 1197 | 31 | 1125 | 25 | 1053 | 17 |

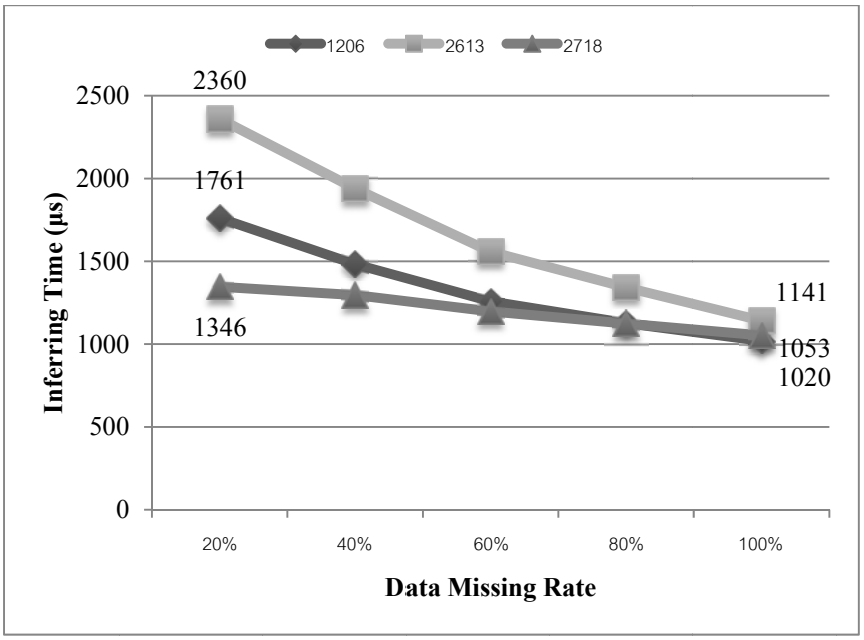



**Figure 5-25: Inferring times of the CATE framework approach at specific missing data rates**

**Table 5-19: Accuracy of the CATE framework approach at specific missing data rates**

| Road Link | Accuracy (%) at specific data missing rate (%) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Miss 20% | | Miss 40% | | Miss 60% | | Miss 80% | | Miss 100% | |
| | Mean | S.D. | Mean | S.D. | Mean | S.D. | Mean | S.D. | Mean | S.D. |
| 1206 | 88.06 | 0.36 | 88.43 | 0.20 | 88.47 | 0.11 | 88.51 | 0.13 | 88.43 | 0.00 |
| 2613 | 81.38 | 0.35 | 81.25 | 0.22 | 81.34 | 0.16 | 81.30 | 0.21 | 80.04 | 0.00 |
| 2718 | 96.59 | 0.16 | 94.75 | 0.15 | 93.13 | 0.14 | 91.54 | 0.14 | 89.47 | 0.00 |

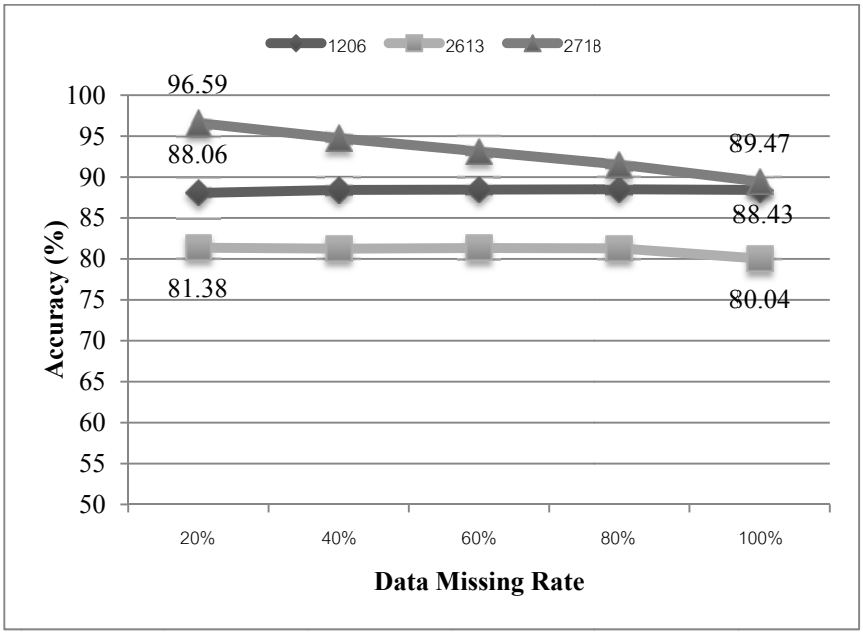



**Figure 5-26: Accuracy of the CATE framework approach at specific missing data rates**

Figure 5-25 indicates that the time taken to infer the missing sensory traffic data of the observed road segments (1206, 2613 and 2718) drops when the missing data rate is higher. This reduction is caused by the reduced size of the selected inference model which is less complicated than the model created for the situation that more context attributes are available. While the accuracy also drops slightly when the missing data rate is higher, it is still over 80% even when the missing rate is 100% (as can be seen in Figure 5-26).

For the road segment 1206 the accuracy does not drop and stays constant even though the missing data rate is higher. This shows that the missing context attributes have a

minimal impact on the traffic conditions of road segment 1206 except for day and time (because day and time are always be available for all missing data rates).

Because our CATE framework is designed to be applicable to all road segments in Bangkok, it is not feasible to consider suitable context attributes specifically for each segment because of the number of total road segments in Bangkok (over 5,241 road segments [130]) and the associated cost of the task. Nonetheless, even though the selected context attributes used to build the inference models for road segment 1206 do not have much impact on the reported traffic condition of road segment 1206, the accuracy is still over 88% and does not drop when the missing data rate is increased.

**Table 5-20: Accuracy of the CATE framework approach at specific missing data rates (*active period*)**

| Road Link | Active period accuracy (%) at specific missing data rates (%) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Miss 20% | | Miss 40% | | Miss 60% | | Miss 80% | | Miss 100% | |
| | Mean | S.D. | Mean | S.D. | Mean | S.D. | Mean | S.D. | Mean | S.D. |
| 1206 | 80.53 | 0.75 | 80.94 | 0.30 | 80.99 | 0.17 | 81.02 | 0.21 | 80.62 | 0.00 |
| 2613 | 63.94 | 0.68 | 63.70 | 0.44 | 63.55 | 0.22 | 63.55 | 0.33 | 61.65 | 0.00 |
| 2718 | 93.65 | 0.36 | 90.15 | 0.33 | 87.19 | 0.29 | 84.12 | 0.27 | 80.06 | 0.00 |

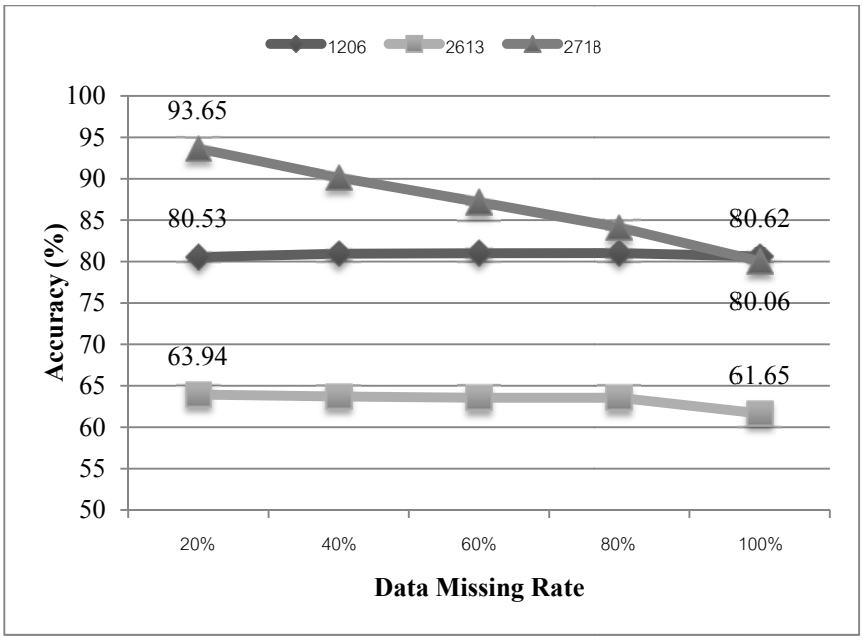



**Figure 5-27: Accuracy of the CATE framework approach at specific missing data rates (*active period*)**

The traffic congestion degree in the active period is more difficult to infer than in the non-active period because the traffic congestion degree in this period is dynamic. This same dynamism also increases the value of information from this period compared to information from a non-active period. In addition, the non-active period is the time period that road users usually do not use the roads, thus pay less attention to the traffic information in the non-active period. Figure 5-27 shows that accuracy falls when the missing data rate is increased (except for road segment 1206) which can be explained by the same reason as in Figure 5-26. However, even the lowest accuracy

obtained using the CATE framework is still higher than the accuracy in the active period of the single model approach. We will further discuss the comparison of the single model approach and CATE framework approach in the next section.

**Table 5-21: Accuracy of the CATE framework approach at specific missing data rates (*non-active period*)**

| Road Link | Non-Active period accuracy (%) at specific data missing rate (%) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Miss 20% | | Miss 40% | | Miss 60% | | Miss 80% | | Miss 100% | |
| | Mean | S.D. | Mean | S.D. | Mean | S.D. | Mean | S.D. | Mean | S.D. |
| 1206 | 95.47 | 0.30 | 95.76 | 0.15 | 95.84 | 0.16 | 95.88 | 0.10 | 96.11 | 0.00 |
| 2613 | 98.43 | 0.13 | 98.65 | 0.05 | 98.79 | 0.09 | 98.78 | 0.08 | 98.12 | 0.00 |
| 2718 | 99.50 | 0.11 | 99.29 | 0.06 | 98.98 | 0.04 | 98.85 | 0.06 | 98.73 | 0.00 |

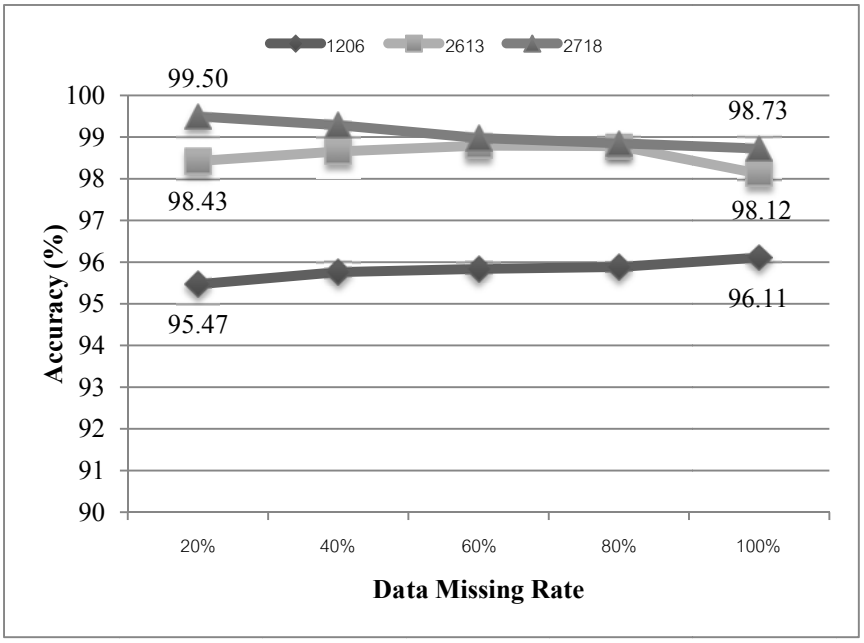



**Figure 5-28: Accuracy of the CATE framework approach at specific missing data rates (*non-active period*)**

Figure 5-28 shows that the accuracy is very high for all road segments when considering only the non-active period. A non-active period is simpler and easier to infer than an active period because the traffic congestion degrees in non-active periods are mostly *L*.

## 5.5 Overall Results Discussion

In this section we compare the results from the three approaches and demonstrate the performance strengths of the CATE framework. We begin by examining the time taken to create the inference data (the inferring time), and then go on to examine the accuracy of the results of each approach.

Figure 5-29 compares the inferring time of the single model approach and the CATE framework. The mode approach is not included in this comparison. Although the inferring time of the mode approach is less than 0.2 μs (because of the simplicity associated with substituting a mode value), the resulting accuracy is not high. As discussed in Section 5.4.1, the mode value is not suitable for use especially in active periods when there is high dependence by drivers on traffic data. Consequently, the mode approach is omitted from this part of the comparison.

The inferring time of the single model approach is unchanged for all missing data rates. This is due to the size of the inference model (the decision tree) remaining constant because the single model approach uses only one inference model for all cases. In contrast, the inferring time of the CATE framework decreases when the missing data rate increases. As the number of missing context attribute increases, the program chooses models that are built for fewer context attributes. This means the size of the inference model (that is, the size of the decision tree) is smaller, thus requiring less inferring time. Even when the missing data rate is as low as 20%, although the inferring time of the CATE framework is more than the inferring time of the single model approach, the difference is only 0.3 milliseconds. This is a negligible amount when compared to the increase of accuracy as shown in Figure 5-30.

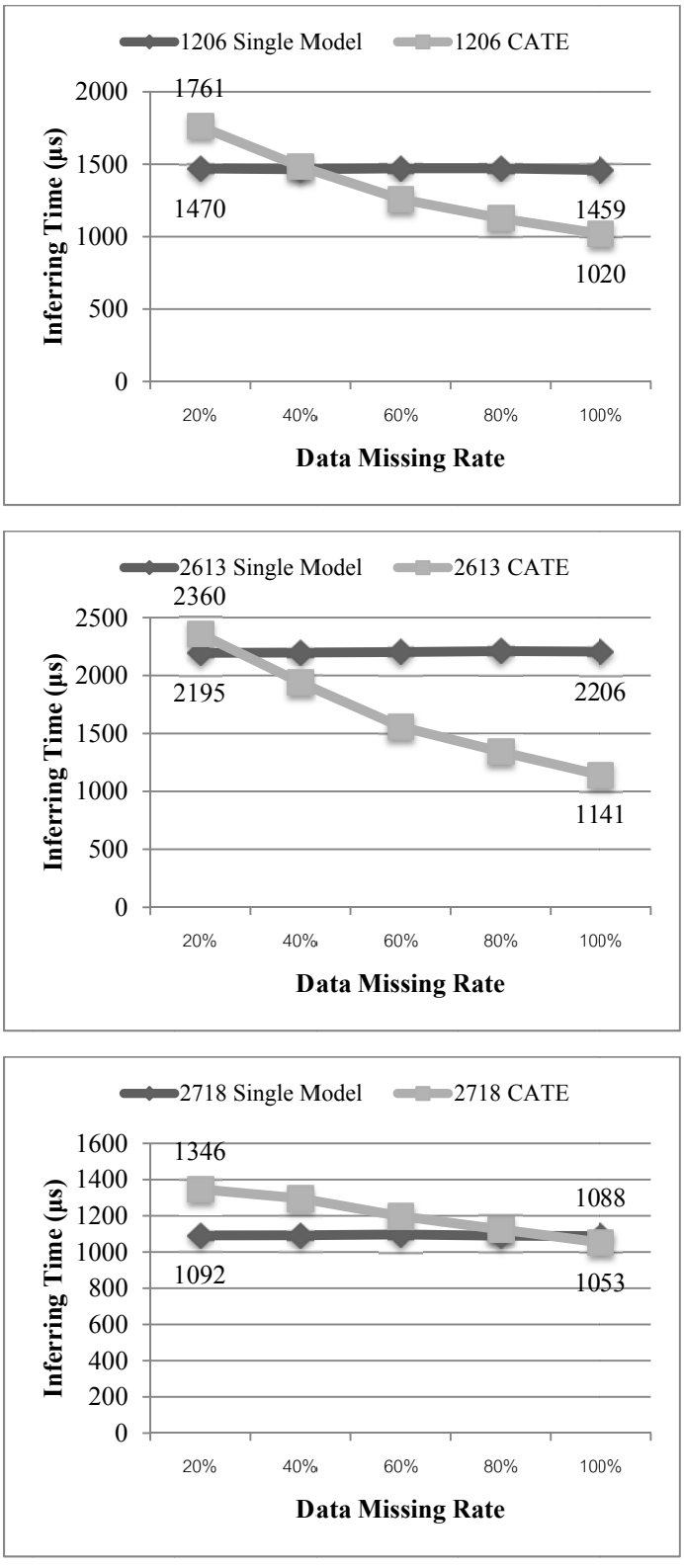



**Figure 5-29: Inferring times of the single model and CATE framework approaches at specific missing data rates for road links 1206, 2613 and 2718**

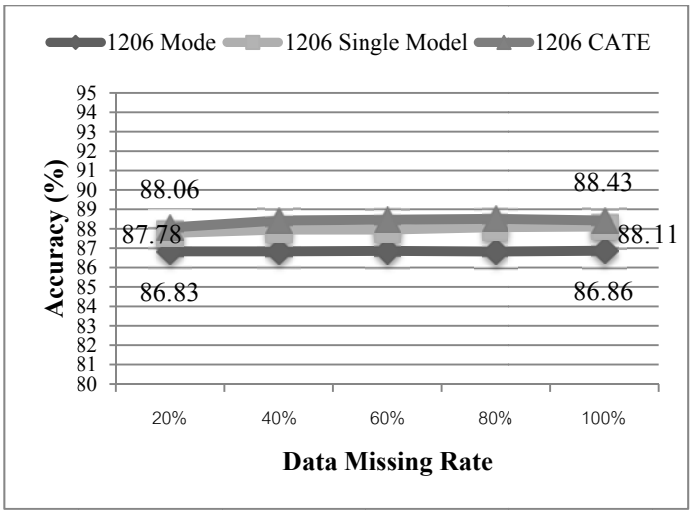





**Figure 5-30: Accuracy of the single model and CATE framework approaches at specific missing data rates for road links 1206, 2613 and 2718**

The graphs in Figure 5-30 demonstrate that the CATE framework performs better than the mode and single model approaches in terms of accuracy. The accuracy of the mode approach is constant for all missing data rates because it only uses a mode value when the program detects missing sensory traffic data for an observed road segment. However, the accuracy of the mode approach is much lower than the single model approach and of our proposed framework.

The CATE framework gives higher accuracy than other solutions for all missing data rates. Even though the accuracy drops when the missing data rate is increased, it is still above 80%. Our CATE framework is resilient to missing rates as high as 100%, or under situations when only day and time context attributes are available.

Because the CATE framework avoids missing context values in inference models by using different inference models for different available real time context attribute sets, higher accuracy can be achieved. In contrast, under the single model approach, the missing context attribute(s) decreases accuracy because of the missing value manipulation. Even though the DBI method that handles missing attribute values of C4.5 can perform better than many other algorithms [136], in regards to accuracy, the CATE framework produces better results than the single model approach.

The superior performance of the CATE framework becomes more explicit when comparing accuracy in an active period, as illustrated in Figure 5-31.

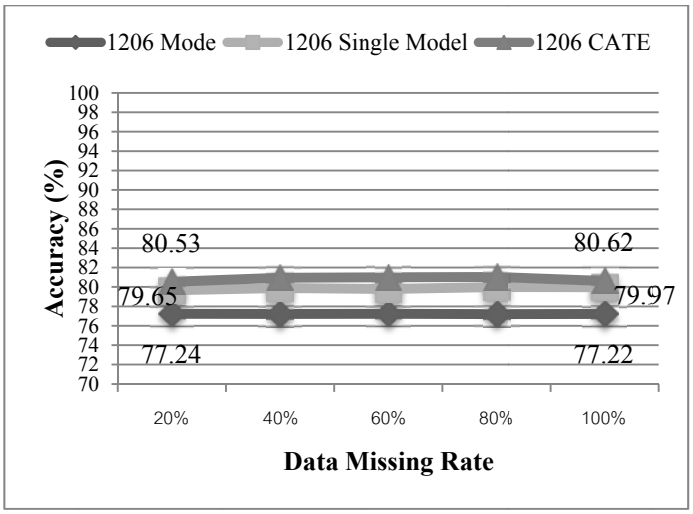





**Figure 5-31: Accurancy of the mode, single model, and CATE framework approaches at specific missing data rates for road segments 1206, 2613 and 2718 (*active period*)**

Figure 5-31 demonstrates that when the missing data rate is increased, the CATE framework gives higher accuracy than the single model approach and outperforms the mode approach. For example, for road segment 2613, our approach, when the missing data rate is 100%, is 10% more accurate than the single model approach.

As we stated previously, the traffic congestion degree in an active period is more difficult to infer than the traffic congestion degree in a non-active period because it is more dynamic. In addition, information about an active period is more useful to road users than information about a non-active period. Therefore, the accuracy is very important in this period. An approach's performance, in terms of its accuracy and speed, in the more difficult active period can therefore be considered more important than its performance in a non-active period. The evidence from our tests suggests that the proposed CATE framework would perform well in providing useful traffic information to road users under the circumstances of an active period.

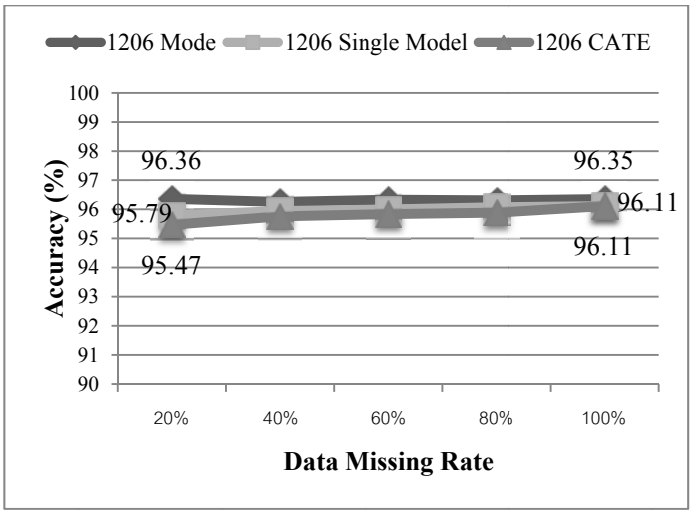





**Figure 5-32: Accuracy of the mode, single model, and CATE framework approaches at specific missing data rates for road segments 1206, 2613, and 2718 (*non-active period*)**

Even though the accuracy of three approaches is approximately the same in the non-active period (illustrated in Figure 5-32), this period is not important because the traffic condition remains a reasonably consistent *L*. In addition, during the non-active period, not many vehicles use the roads and there is not the imperative for users to know the traffic information.

The results of our tests demonstrate the feasibility and the efficiency of our approach. It can be seen that even when only one acquirable context attribute is applied, the accuracy rate is over 70%. Moreover, when a greater number of influential context attributes are involved, in most cases, the accuracy rate is above 80%. The evaluation results also prove that our CATE framework is resilient to high missing data rates even though no other context attributes may be available (apart from day and time, which are always acquirable).

The CATE framework can be used even when the collected historical data is sparse. Our evaluation shows that the CATE framework produces accurate results even though the historical data on which the learning and model building is based may have accumulated only one month's worth of data, or about 10,000 records. The system also becomes smarter with additional new actual data because the framework has relearning capabilities and rebuilds models at pre-defined time periods. Our framework could thus also be applied to road segments that have yet to establish historical data such as road segments with newly installed traffic sensors.

In general, it can be speculated that adding more contexts to a model should improve accuracy. However, we found that adding too many contexts can yield greater processing time required and sometimes even reduce accuracy, as evidenced in the results. In addition, it is noticeable from the results that the combination of the day and time context attributes contributes to high accuracy, and, when considered along with other context attributes, this combination helps to improve accuracy substantially. Furthermore, the day and time context attributes are considered known facts that can be obtained with without or little cost and effort.

Our conclusion in relation to the day and time context attributes is that temporal contexts, whenever available, should always be included as context attributes. Following this realisation, and looking back at Table 5-3 that listed the potential

context sets, the number of inference models to be built could perhaps be reduced to eight (CtxSet15, CtxSet23, CtxSet24, CtxSet25, CtxSet28, CtxSet29, CtxSet30 and CtxSet31). This leads to the possibility that we could substantially reduce the number of inference models by building only inference models for the sets that include context attributes (such as day and time) that have a significant impact on road traffic estimation. In turn, reducing the number of inference models would reduce the amount of processing time required, again minimising cost and effort. Further consideration is given to this possibility in Chapter 6 where we investigate the number and type of influential context attributes in order to reduce the number of inference models.

## *5.6 Chapter Summary*

The CATE framework is designed to provide an inferred traffic congestion degree from acquirable context attributes when sensory traffic data is unavailable. This chapter presents the details of Evaluation I, which aims to evaluate the feasibility of the CATE framework and its accuracy in comparison to two other solutions: the mode value and the single model approaches. In most cases, the accuracy of the CATE framework is above 80% at different missing data rates. The proposed framework also gives higher accuracy than the mode value and the single model approaches.

The results prove that our CATE framework is robust, with high estimation accuracy, even when subject to high levels of missing context attributes. Although our approach requires more model building time in order to build inference models than the single model approach, the amount of time required is negligible for the current generation of computer technology. The evaluation demonstrates the feasibility and efficiency of our CATE framework.

We also found that adding too many contexts can yield more processing time required and sometimes even reduce accuracy. In addition, we proposed that the attributes day and time, whenever available, should always be included because they help improve accuracy when used together with other attributes and can usually be obtained with minimal cost and effort. Following this line of thought, we proposed that we can substantially reduce the number of inference models by only building inference models from context attribute sets that include both day and time. Further investigation to improve the proposed framework by determining the most appropriate

constituents of a context attribute set, followed by an evaluation, is conducted in the next chapter.

# Chapter 6 Refinement of the CATE Framework

Evaluation I in Chapter 5 demonstrates the feasibility of the CATE framework. The framework performs well, is resilient to high missing data rates and gives better accuracy than other approaches. However, we observed that performance could be improved by reducing the time taken to build the models. Reduced time can be achieved through the careful selection of relevant context attribute sets, as not all combinations of context attributes will aid in good estimating the missing sensory data, and some context attributes would require too much effort to obtain them. The reduced number of the set of context attribute combination, in turn, will reduce the total number of inference models that needs to be built. The reduction in the total number of inference models would also lead to the faster location of the appropriate inference model during deployment. Reducing the number of context attributes required would also reduce the cost of system implementation (in terms of required computing resources, human resources, effort and time).

To select the reduced set of context attributes, we noted the result of our evaluation of the CATE framework discussed in Chapter 5. We then used an online survey to cross-reference this selection with the perceptions of Bangkok road users on this issue. Finally, we evaluated the performance of the reduced set in terms of accuracy and processing time and compared the performance of the reduced set against the performance of the full set described in Chapter 5. This knowledge is used to improve our designed artefact and to yield the final artefact. The refinement of the CATE framework and the evaluation of its performance are presented in this chapter.

The purpose of the evaluation described in this chapter, referred to as Evaluation II, is to assess the improvement of the refined artefact in terms of accuracy and processing time. The impact on required resources and cost are also discussed. In this chapter we also evaluate our refined CATE framework against other approaches.

The chapter begins with a statistical analysis of the factors influencing Bangkok's traffic drawn from our web-based survey. Next, an analysis is presented relating to the reduced set of influential context attributes. These were developed based on the results from Evaluation I and the survey results concerning the perceptions of Bangkok road users. We make an adjustment to the initial artefact and present the final version of the artefact: the refined CATE framework. This is followed by an evaluation of the improvement of the refined artefact. To finish this chapter, an implementation of the refined CATE framework is explained to demonstrate the feasibility of the framework.

## *6.1 An investigation of the Factors Influencing Bangkok Traffic Conditions in the Perceptions of Bangkok Road Users*

The context attributes used in the CATE framework were initially chosen based on the literature review and the suggestions of researchers expert in ITS. Even though the evaluation of our proposed framework gave promising results, we also investigated further by studying the influential factors (IF) that affect traffic congestion in Bangkok based on road user perception. This was to confirm the feasibility of our selected influential context attributes explained in Chapter 4 and Chapter 5, and to help confirm the degree of importance of each influential context attribute. This confirmation, combined with the knowledge gained from Evaluation I, was used to guide the final selection of the influential context attributes set, or what we call the reduced set of influential context attributes (*RS*). Our research was carried out based on a quantitative web-based survey. The survey overview, survey methodology and process, and the analysis of the result are described in the following sections.

### 6.1.1 Survey Overview

The research methodology adopted in this study is quantitative research via a web-based survey to collect answers from respondents. The terms "Internet", "web", and "online" survey have been used in research interchangeably to refer to any survey in which data is collected via the Internet [137]. Online surveys were first utilized in 1995 and have since gained popularity. An online survey can take place in different ways such as through execution on respondents' machines (client-sided) or execution on the survey organizations' web servers (server-sided) [138].

The major reasons for using an online survey over a traditional paper-based survey are reduced costs, faster response times and higher quality data because incomplete and erroneous data can be prevented [139-141].

In this study, the server-sided "surveymonkey.com" website was used. Surveymonkey.com hosts surveys on its website and respondents are directed to a web-page containing questionnaires. Respondents can participate in a study by submitting their answers to the server by clicking a "Next" or "Finish" button. Because of issues raised in [142], an "intercept survey" was selected as the sampling method for this study. The intercept-based approach is a probability-based method for online surveys that target the visitors of particular websites.

The questionnaire of our survey consisted of six parts: *Screening Questions*, *Demographic Information*, *Bangkok Traffic Questions*, *Traffic Information Consumption*, *Traffic Awareness in Social Networks* and *Traffic Content in Social Networks*. To ensure accurate and useful answers, the screening questions were designed to ensure that only individuals who drive regularly in Bangkok proceeded to the rest of the survey. The method of designing and distributing this online survey is described in the next section.

### 6.1.2 Survey Methodology and Process

In this section, the design of our survey and its validation are described. How we deploy the survey after it was well designed and validated and how we analyze data are also explained.

**Survey Design**

The process started with designing the questionnaire using questionnaire design theory [138, 143]. The questionnaire consisted of 20 questions in total, categorized into six parts. The first part, *screening questions*, was used to screen for drivers who drive a car regularly in Bangkok. The second part related to *demographic information*. Demographic information is useful for determining information relating to particular groups of respondents (for example, when cross referencing by age, education or gender). The third part asked about the perceptions of Bangkok road users in relation to Bangkok traffic. The fourth part was about the information needs of Bangkok road

users and their preferred communication channels. This information is useful for designing traffic report services and applications. The fifth part asked about the social network usage of Bangkok road users while the last section asked about traffic information in social networks. These last sections were included in order to study the potential use of social networks for TIS in Bangkok. The survey questions can be found in Appendix C.

**Survey Design Validation**

After the survey was designed, it was pilot tested with a Cronbach's alpha of 0.84 for internal consistency reliability. This indicated the designed questionnaire was valid and ready to distribute.

**Survey Deployment**

After the pilot test, the questionnaire was distributed over the Internet through the surveymonkey.com service. The web link to the questionnaire page was disseminated during December 2012 to February 2013 through websites, web boards and the researcher's own social network. The link to the online questionnaire page was also shared through social networks such as Twitter, Facebook and Google+ via the researcher's own accounts. Because the researcher had over 400,000 followers on her social media sites, it was possible to stimulate a large number of responses – 3,037 – within the short period of three months. This response highlights the success of using social networks to reach out to massive numbers of people and aid in overcoming one of the limitations of surveys – a poor response rate. Any person who volunteered to undertake the survey could click the link and start answering the questions with no obligation.

The first page of the online questionnaire contained an explanation of the questionnaire's purpose, the approximate time to complete the survey and assurances of privacy and confidentiality.

The screening question was then used to determine whether the participant routinely drove a vehicle in Bangkok. If the answer was *no*, the survey was terminated because that participant did not belong to our target group. If the participant answered *yes* to

the screening question, he or she proceeded to answer the rest of the questions. Participants were able to leave the survey at any time.

**Data Analysis Preparation**

After the participants completed and submitted the online questionnaire, their answers were stored electronically. A data file of 3,037 respondents was generated in a SPSS compatible format. Only complete records (records where participants had answered all questions) with no errors and that had passed the screening question were collated for further research. Finally, 861 out of 3,037 records remained for analysis.

The next step was to determine whether this sample size was sufficient for analysis to give valid results. This was achieved by calculating the population size for our research. According the number of driving licenses issued under the Motor Vehicle Act and reported by the Department of Land Transport of Thailand [144], in 2012, the number of driver licenses issued in Bangkok was 5,777,806. This number can be considered the population size ($N$) for our research. According to the sample size determination for research activities in [145], our sample size $S$ = 861, which is the number of respondents who drive routinely in Bangkok and had completed the survey, is sufficient. Responses from participants were then transformed into code and analysed by SPSS software version 17.0. The results are described in the following section.

**Data Analysis**

The analysed results were categorized into five groups: demographic information, perception of Bangkok traffic, traffic information consumption, factors influencing Bangkok traffic conditions and the potential use of social networks for TIS. In this chapter, only the parts of the study that relate to factors influencing Bangkok traffic conditions are presented. The demographic information of respondents is also reported in this chapter. The remaining survey results and analysis, which can be used to help guide TIS implementation, can be found in Chapter 7.

### 6.1.3 The Demographic Information of Respondents

According to the data, 77.8% of respondents were male while 22.2% were female. The majority of respondents were aged between 26 and 41 years old. The top two education levels of the respondents were a Bachelor's degree (61%) and a Master's degree (26.8%), respectively.

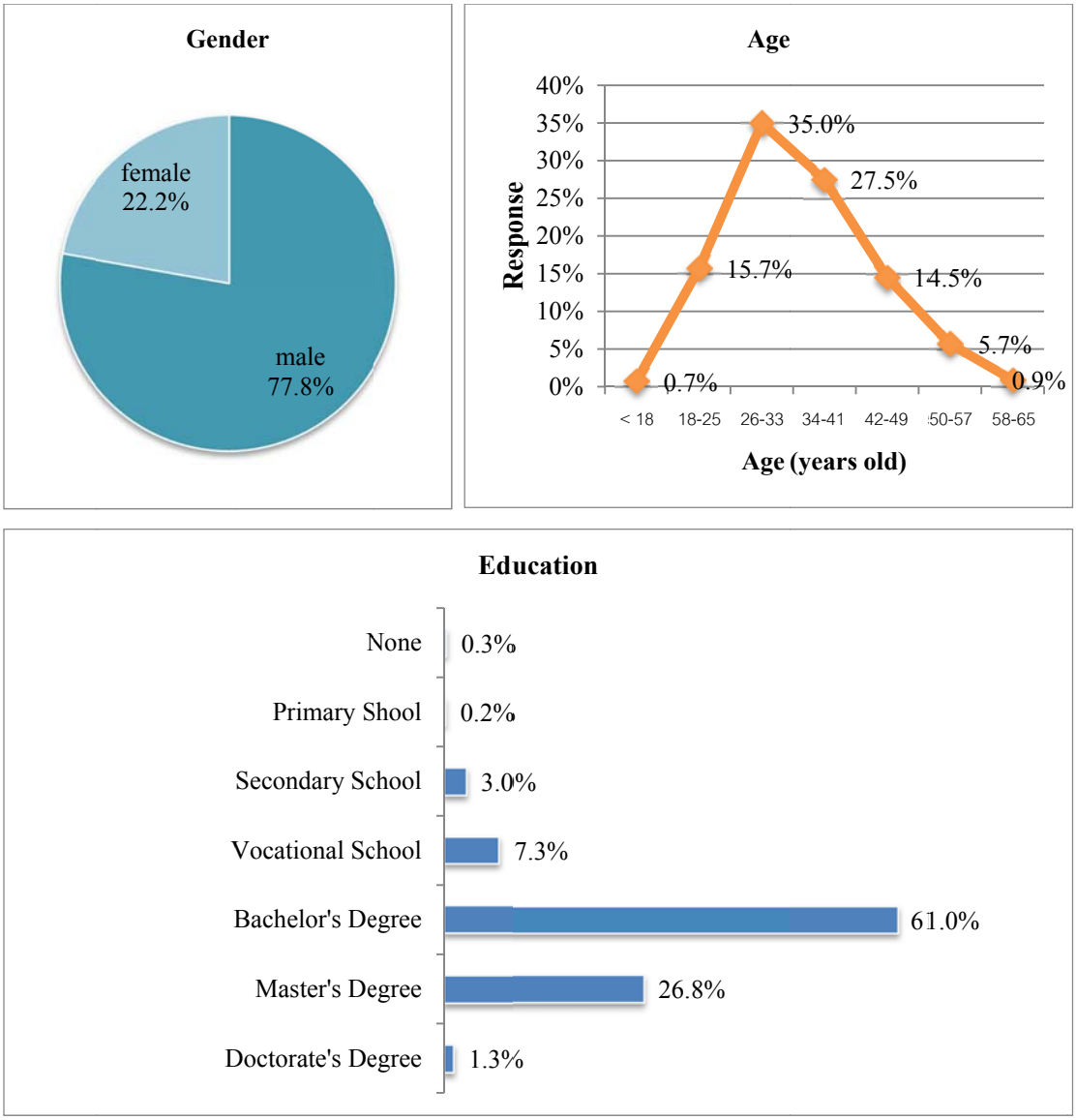



**Figure 6-1: Gender, age, and education of respondents**

Most respondents were unmarried (60.9%) while 73.1% were employed. Of these, 49.1% were employees of a private company and 24% were entrepreneurs.

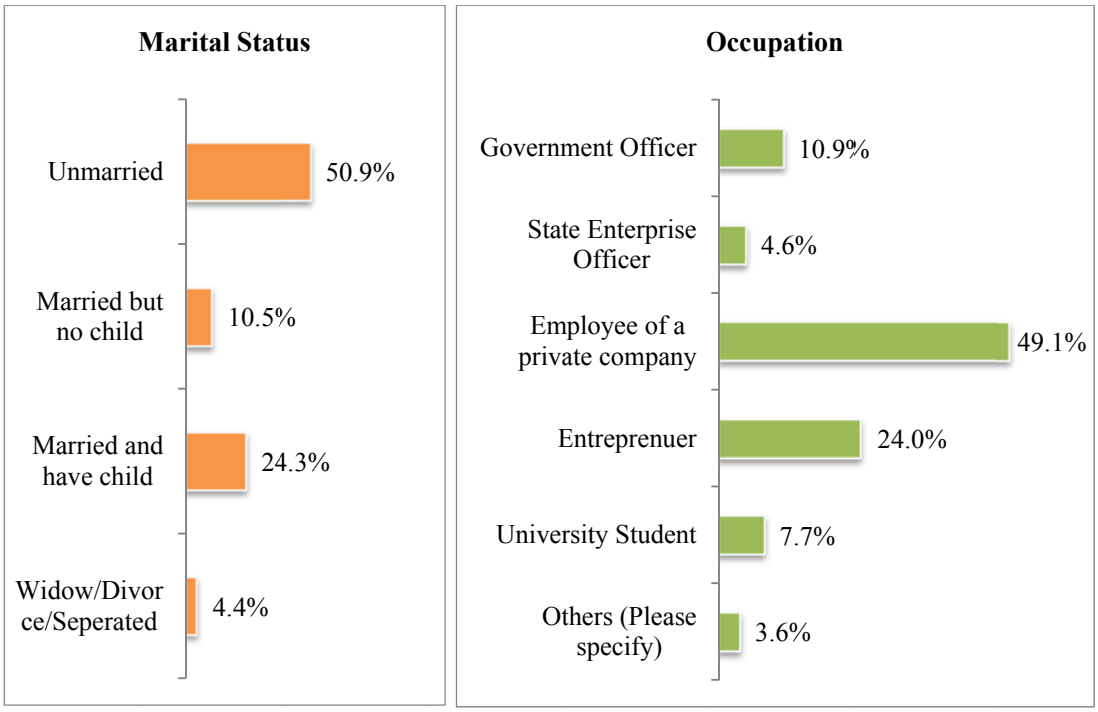



**Figure 6-2: Marital status (left) and occupation of respondents (right)**

About half of the respondents (49.5%) drive vehicles every day. The survey results report that 92.6% of respondents are cars owners. It is noticeable that about 70% of respondents live outside the central business district. This could be because of the screening question that ensured only individuals who drive cars routinely completed the survey. As of 2012 – 2013, no sky train (BTS) or underground train (MRT) services reach the area outside the city of Bangkok. Consequently, most residents have their own cars.

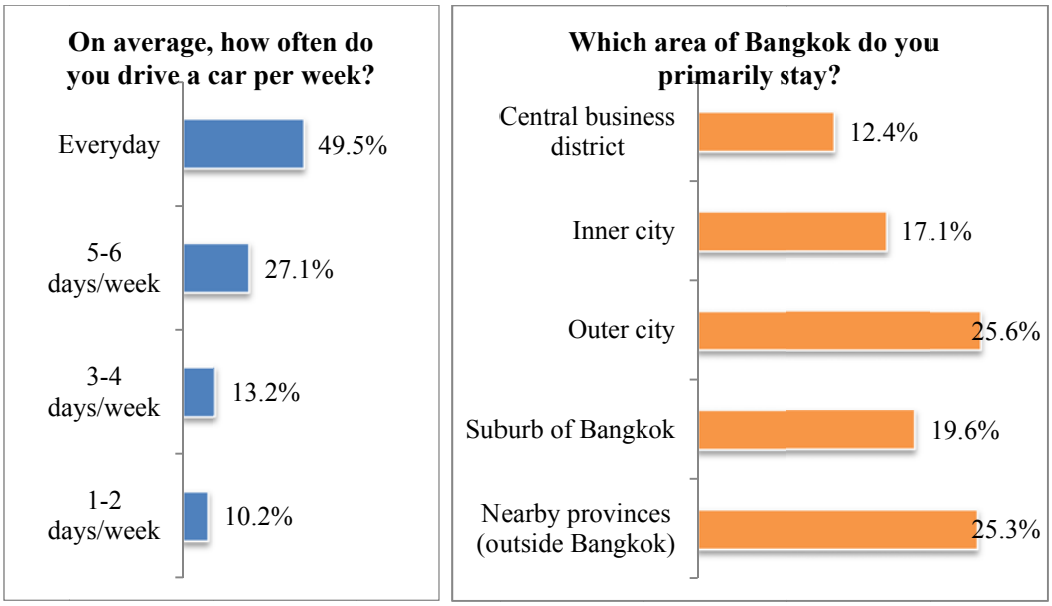



**Figure 6-3: The frequency of driving cars (left) and living areas of respondents (right)**

Figure 6-4 illustrates respondents' main reasons for driving. Respondents gave a score from 1 to 5 for each reason where 1 represented "Never" and 5 represented "This is the main reason". The average score shows that "Work/Study" (4.39) and "General Convenience" (3.96) were the main reasons respondents gave for driving.

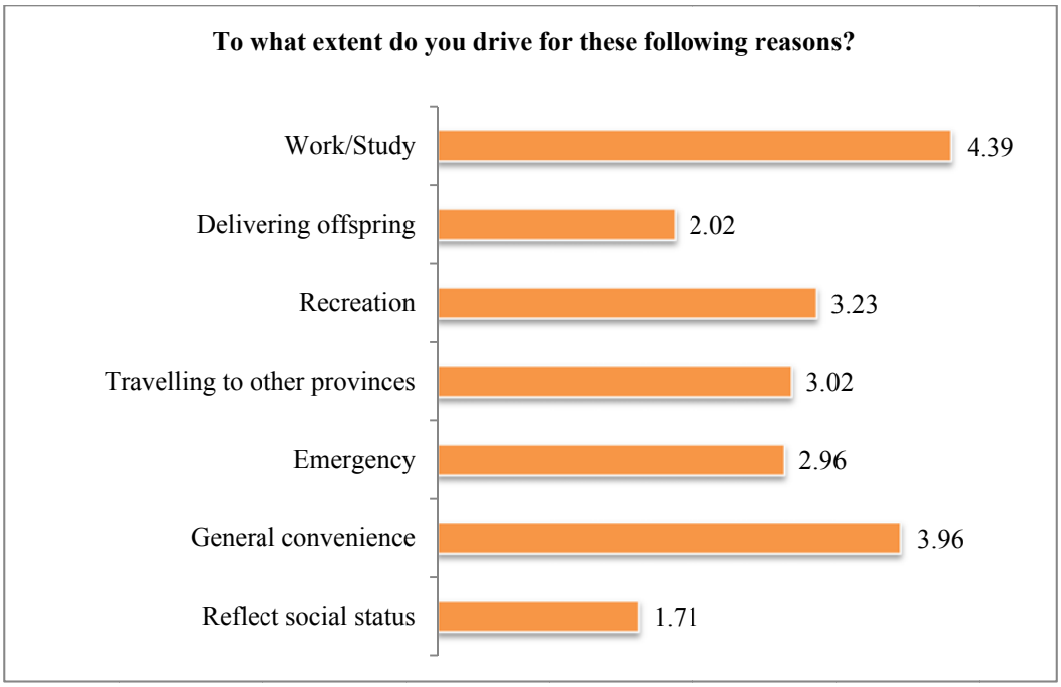



**Figure 6-4: Respondents' reasons for driving**

## *6.2 Traffic Information Needs*

Figure 6-5 illustrates the traffic information needs of Bangkok drivers. We found that 95% of drivers in Bangkok are aware of the usefulness of knowing traffic information. A large proportion - 38% - thinks that it is "Very useful" to know traffic information while 40.2 % found it "Useful". Only 16.7 % of drivers think that it is of little use to know traffic information. This demonstrates that the majority of Bangkok drivers value traffic information. It is logical to assume that a TIS that can report traffic conditions to drivers would thus also be considered useful.

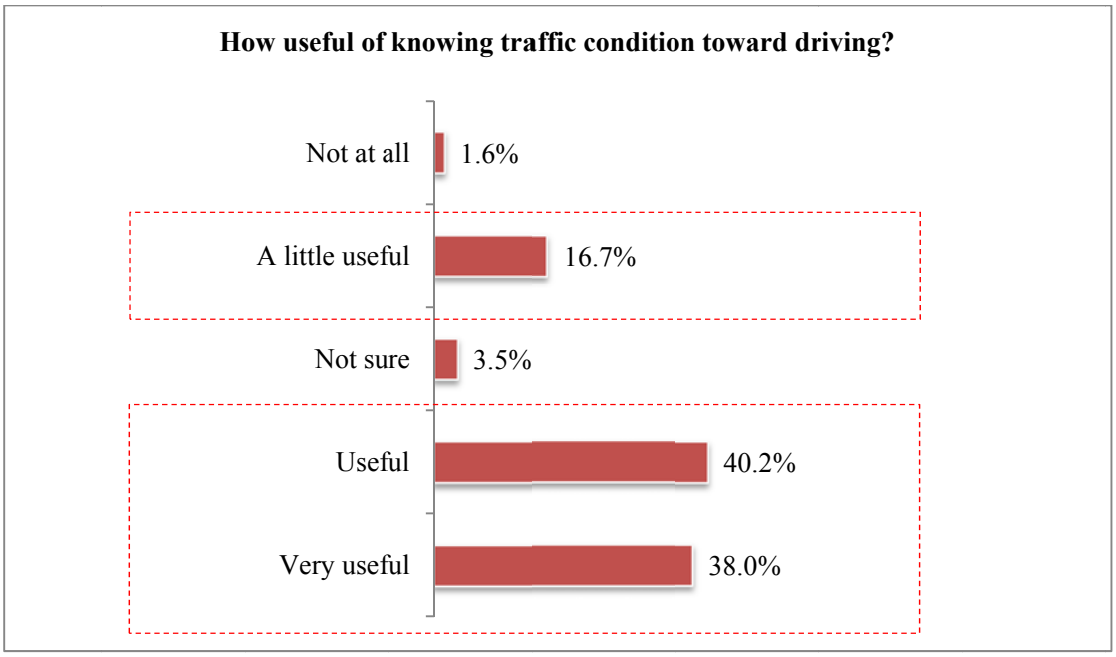



**Figure 6-5: Traffic information needs of respondents**

From the results of questions regarding the traffic information needs of Bangkok drivers, we can conclude that most of the road users in Bangkok want to know traffic information. This confirms the usefulness of knowing traffic information of Bangkok road users and shows that our proposed framework, through its ability to inform TIS, is useful. In addition, any provision of a traffic information dissemination system, whether supplied through government sectors or private organizations, would be of value to Bangkok road users.

## *6.3 Respondents' Perceptions of Bangkok Traffic*

When respondents were asked to choose a description of Bangkok traffic, most drivers selected "Traffic Jam" (78%). The second rank was "Too much car density on roads" (11.6%) while "Lack of Discipline" ranked third (7.2%).

These results, illustrated in Figure 6-6, suggest that in the perception of road users, traffic jams are a critical problem in Bangkok.

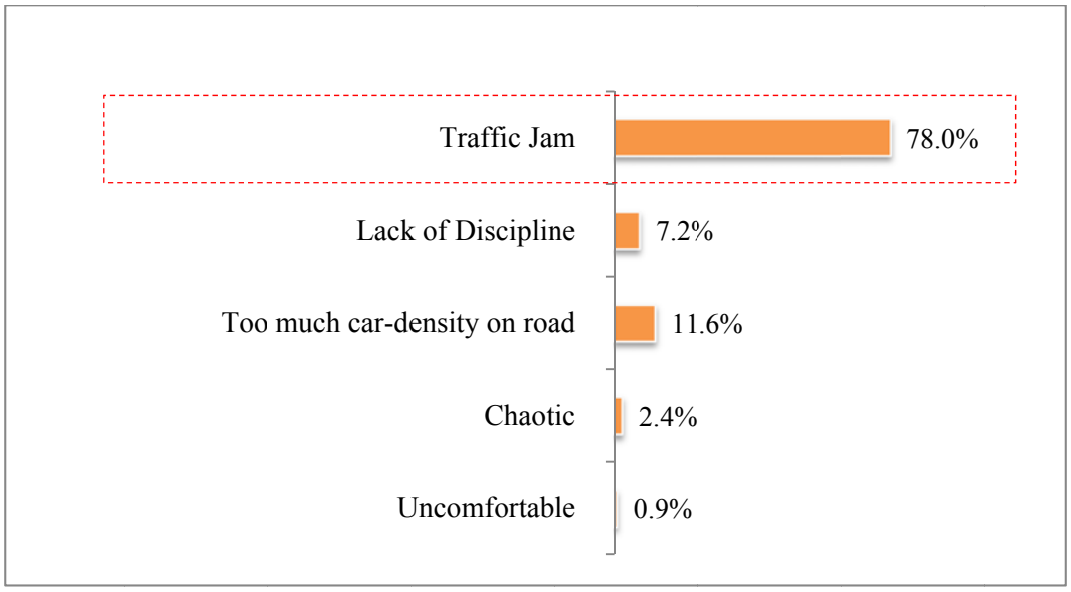



**Figure 6-6: Respondents' perception of Bangkok traffic**

### 6.3.1 Factors Influencing Bangkok's Road Traffic Conditions

In order to discover the factors that influence Bangkok's traffic in the perception of road users, in the survey we asked road users the question "To what extent do you think the following factors affect the traffic in Bangkok?" The respondents give a score ranking from 1 to 5 where 1 represented "No effect" and 5 represented "Extreme effect". The average score of each item was calculated and sorted in ascending order.

The result shows that the "Specific time of the day" and "Incidents" have the highest ranking follow by "Level of rainfall", "Density of vehicles on the road adjacent to the observed road", "Public holiday", "Specific day of the week" and "Group of the days" respectively as illustrated in Figure 6-7.

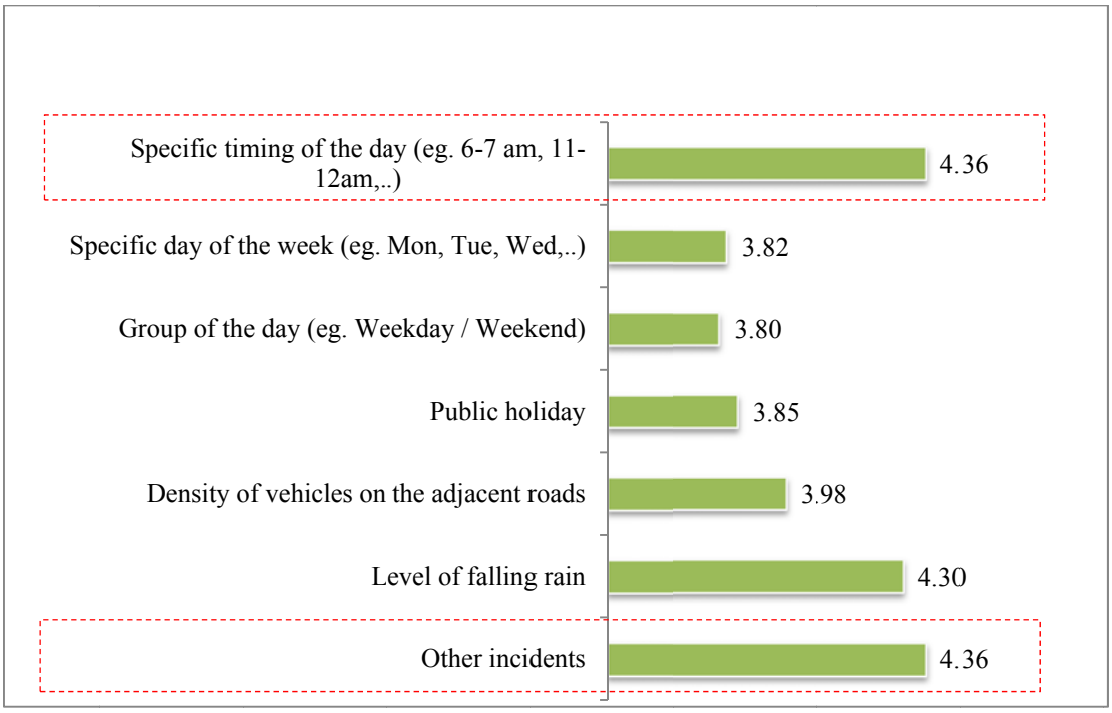



**Figure 6-7: Road users' perceptions of factors influencing Bangkok traffic**

In order to be able to specify the level of importance or the weight of each factor in the perception of Bangkok road users, we use the class interval technique. The seven factors are "Specific timing of the day", "Specific day of the week", "Group of the day", "Public holiday", "Density of vehicles on the adjacent roads", "Rain level", and "Other incidents" (such as accidents and road construction). The level of importance ranges from "1" for no effect, "2" for slight effect, "3" for medium effect, "4" for high effect and "5" for extreme effect.

The Weight Mean Score (WMS) method below from [146] was applied.

$$WMS = \frac{5F_5 + 4F_4 + 3F_3 + 2F_2 + 1F_1}{TNR}$$

where *WMS* = Weight Mean Score and

$F_5$ = the frequency count of "extreme effect",

$F_4$ = the frequency count of "high effect",

$F_3$ = the frequency count of "medium effect",

$F_2$ = the frequency count of "slight effect",

$F_1$ = the frequency count of "no effect"

and *TNR* = Total number of respondents.

The class interval calculation was based on the following formula where *k* is the number of classes).

$$\text{Class interval} = \frac{x_{max} - x_{min}}{k} = \frac{5-1}{5} = 0.80$$

Consequently, the criteria for interpreting the findings were 4.21-5.00 for extreme effect; 3.41-4.20 for high effect; 2.61-3.4 for medium effect; 1.81-2.60 for slight effect and 1.00-1.80 for no effect.

**Table 6-1: Descriptive statistics of IF analysis**

| | N | WMS | S.D. | Interpretation |
|---|---|---|---|---|
| Specific timing of the day (eg. 6-7 am, 11-12am, ..) | 861 | **4.36** | 0.846 | Extreme effect |
| Specific day of the week (eg. Mon, Tue, Wed, ..) | 861 | 3.82 | 0.969 | High effect |
| Group of the day (eg. Weekday / Weekend) | 861 | 3.80 | 1.004 | High effect |
| Public holiday | 861 | 3.85 | 1.141 | High effect |
| Density of vehicles on the adjacent roads | 861 | **3.98** | 0.958 | High effect |
| Rain Level | 861 | **4.30** | 0.845 | Extreme effect |
| Other incidents (eg. Road construction, accident, protest rally) | 861 | 4.36 | 0.857 | Extreme effect |

From the results shown in Table 6-1, we can group the seven factors into two groups. Group 1 incorporates the factors that, in the opinion of road users, have a high effect on Bangkok road traffic: *day of the week, group of the day, public holiday* and *the density of vehicles on the adjacent roads*. Group 2 comprises those factors that have an extreme effect on Bangkok road traffic: *time, rain level* and *incidents.*

The survey results concerning the traffic information needs of Bangkok road users confirm the usefulness of knowing traffic conditions in Bangkok. The survey results and analysis of factors influencing Bangkok's road traffic conditions confirmed our initial selection of influential context attributes. We then proceeded with further analysis in order to reduce the set of influential context attributes. This analysis is explained in the next section.

## 6.4 Analysis to Produce a Reduced Set of Influential Context Attributes

To reduce the total time to learn and build inference models, we can reduce the number of models to be built by reducing the number of influential context attributes applied; the accuracy, however, must remain high. We call the set of selected influential context attributes a *reduced set* of influential context attributes or *RS*. Reducing the number of influential context attributes leads to a decreased number of inference models being built. This then reduces the total processing time to learn and build the inference models, the cost of acquiring context attributes and the amount of data preparation required, thus improving the overall performance of the deployment of framework.

In order to select the context attributes to be included in *RS*, we analyze the experiment results from Chapter 5. The *RS* is the minimal set of influential context attributes that can retain accuracy and improve overall performance. From the results in Table 5-15, Table 5-16, and Table 5-17, it is noticeable that applying only *day* or only *time* gives lower accuracy than applying both. The combination of the *day* and *time* context attributes gives a higher accuracy rate and helps substantially improve the accuracy of results when combined with other context attributes. Moreover, *day* and *time* are considered known facts that can usually be obtained with minimal or no cost. Furthermore, whenever we obtain *time*, we can always obtain *day* from the system clock and vice versa. Hence, both *day* and *time* should always be included in a context attribute set.

Furthermore, the results shown in Table 5-15, Table 5-16 and Table 5-17 in Chapter 5 demonstrate that applying *school* to build the inference models does not significantly

help improving the accuracy rate. On the contrary, including *school* increases the total time to build all inference models and increases pre-processing time. For these reasons, we decided to exclude the *school* context attribute from our reduced sets. The selected influential context attributes are now reduced to *day*, *time*, *rain* and *traffic congestion degree of connected road segment(s)*. To make it more convenient to read, we select only the context attribute sets comprising *day*, *time*, *rain* and *traffic congestion degree of connected road segment(s)* from Table 5-15, Table 5-16, and Table 5-17 to show in Table 6-2, Table 6-3 and Table 6-4 respectively.

Excluding the context attributes sets that do not include the combination of the *day* and *time* attributes, and also excluding those sets with the *school* context attribute, reduces the maximum number of inference models to be built from sixty three models (see Table 5-15, Table 5-16, and Table 5-17) to only eight models (see Table 6-2, Table 6-3, and Table 6-4).

As a consequence of this reduction, the total time for machine learning and inference model building is also reduced considerably while accuracy remains high. These results can be seen in Table 6-2, Table 6-3 and Table 6-4. Table 6-5 illustrates the comparison between the time necessary to learn and build the sixty three models (when applying all context attributes) and the time necessary to learn and build the eight models (when applying only five context attributes).

**Table 6-2: Accuracy of model evaluation and model building time for road link 1206 when using 5 context attributes**

| CtxSet | Context Attributes Applied | Accuracy (%) | | Model Building Time (µs) | |
|---|---|---|---|---|---|
| | | Mean | S.D. | Mean | S.D. |
| 7 | {Day,Time} | 90.61 | 0.11 | 4147 | 338 |
| 22 | {Day,Time,Rain} | 90.48 | 0.13 | 5344 | 2196 |
| 24 | {Day,Time,1211} | 92.70 | 0.06 | 4917 | 1192 |
| 25 | {Day,Time,1403} | 92.08 | 0.12 | 4483 | 851 |
| 43 | {Day,Time,Rain,1211} | 92.83 | 0.06 | 5514 | 575 |
| 44 | {Day,Time,Rain,1403} | 92.11 | 0.11 | 4922 | 576 |
| 47 | {Day,Time,1211,1403} | 93.75 | 0.08 | 6136 | 1866 |
| 59 | {Day,Time,Rain,1211,1403} | 93.78 | 0.08 | 6742 | 977 |
| | **Total Model Building Time (µs)** | | | **42205** | |

**Table 6-3: Accuracy of model evaluation and model building time for road link 2613 when using 5 context attributes**

| CtxSet | Context Attributes Applied | Accuracy (%) | | Model Building Time (µs) | |
|---|---|---|---|---|---|
| | | Mean | S.D. | Mean | S.D. |
| 7 | {Day,Time} | 82.49 | 0.10 | 4613 | 559 |
| 22 | {Day,Time,Rain} | 82.42 | 0.12 | 4476 | 593 |
| 24 | {Day,Time,3015} | 85.06 | 0.13 | 5083 | 467 |
| 25 | {Day,Time,2614} | 84.47 | 0.14 | 5351 | 887 |
| 43 | {Day,Time,Rain,3015} | 85.31 | 0.12 | 5860 | 667 |
| 44 | {Day,Time,Rain,2614} | 84.61 | 0.13 | 6034 | 423 |
| 47 | {Day,Time,3015,2614} | 87.23 | 0.13 | 6877 | 357 |
| 59 | {Day,Time,Rain,3015,2614} | 87.17 | 0.12 | 7623 | 525 |
| | **Total Model Building Time (µs)** | | | **45916** | |

**Table 6-4: Accuracy of model evaluation and model building time for road link 2718 when using 5 context attributes**

| CtxSet | Context Attributes Applied | Accuracy (%) | | Model Building Time (µs) | |
|---|---|---|---|---|---|
| | | Mean | S.D. | Mean | S.D. |
| 7 | {Day,Time} | 89.99 | 0.07 | 4453 | 482 |
| 22 | {Day,Time,Rain} | 90.11 | 0.05 | 4533 | 864 |
| 24 | {Day,Time,2515} | 98.85 | 0.02 | 4747 | 815 |
| 25 | {Day,Time,1401} | 90.22 | 0.09 | 4895 | 1821 |
| 43 | {Day,Time,Rain,2515} | 98.87 | 0.02 | 5117 | 662 |
| 44 | {Day,Time,Rain,1401} | 90.60 | 0.09 | 4486 | 372 |
| 47 | {Day,Time,2515,1401} | 98.97 | 0.01 | 4895 | 550 |
| 59 | {Day,Time,Rain,2515,1401} | 98.99 | 0.01 | 5896 | 1447 |
| | **Total Model Building Time (µs)** | | | **39024** | |

It is noticeable that the graphs of accuracy within Table 5-17 and Table 6-4 indicate that the traffic congestion degree of the connected road ID *2515* is a dominating factor. It gives extremely high accuracy even though it is applied alone to the inference model. In addition, any context attribute set (CtxSet) that includes *2515* also gives similar accuracy. However, this is a special case, which is not applicable to other road links that we have tested in the experiments. Thus, *traffic congestion degree of connected road links* such as the *2515* alone is not sufficient to compose our reduced sets.

**Table 6-5: Model building time comparison when using all context attributes vs 5 context attributes (taken from model evaluation in Table 5-15, Table 5-16, and Table 5-17)**

| Road Segment ID to be Inferred | Model Building Time (µs) | | Reduction (%) |
|---|---|---|---|
| | All context attributes | 5 context attributes | |
| 1206 | 304678 | 42205 | 86.15% |
| 2613 | 318136 | 45916 | 85.57% |
| 2718 | 299445 | 39024 | 86.97% |

In addition, the descriptive statistic of influential factor (IF) analysis obtained from the survey result illustrated in Table 6-1 confirms our context attributes selection. The Weight Mean Score (WMS) in Table 6-1 shows that the IF that has the most impact on the traffic conditions, in the opinions of road users', opinion are *incidents* and the *time period of within each day* (WMS = 4.36). The second is *rain* (WMS = 4.30) while the third is the *traffic congestion degree of adjacent roads* (WMS = 3.98).

According to the opinions of road users in Bangkok, the WMS of *rain* is higher than that of the *traffic congestion degree of adjacent roads*. However, the experiment results show that adding *rain* to the context attribute sets gives only minimal or no accuracy improvement in some cases, while the time to learn and build the inference models is higher. In addition, to obtain the *rain* context attribute, additional sensors to detect rain volume is required. Installing extra rain sensors increases costs. It might not be worth the investment, especially when the budget is limited. Although it is possible to pull real time rain data from organizations that maintain weather data (such as the Thai Meteorological Department), the acquisition of real time rain data is complicated. The system must be linked to the system of the Thai Meteorological Department, which is a government agency. The process of connecting the system and coordinating across organizations is complex. Moreover, the data access policy may also change. Consequently, we exclude the context attribute set that includes "*rain*" from Table 6-2, Table 6-3 and Table 6-4 and show the remaining sets in Table 6-6, Table 6-7 and Table 6-8.

**Table 6-6: Accuracy of model evaluation and model building time for road link 1206 when using 4 context attributes**

| CtxSet | Context Attributes Applied | Accuracy (%) | | Model Building Time (µs) | |
|---|---|---|---|---|---|
| | | Mean | S.D. | Mean | S.D. |
| 7 | {Day,Time} | 90.61 | 0.11 | 4147 | 338 |
| 24 | {Day,Time,1211} | 92.70 | 0.06 | 4917 | 1192 |
| 25 | {Day,Time,1403} | 92.08 | 0.12 | 4483 | 851 |
| 47 | {Day,Time,1211,1403} | 93.75 | 0.08 | 6136 | 1866 |
| | **Total Model Building Time (µs)** | | | **19682** | |

**Table 6-7: Accuracy of model evaluation and model building time for road link 2613 when using 4 context attributes**

| CtxSet | Context Attributes Applied | Accuracy (%) | | Model Building Time (µs) | |
|---|---|---|---|---|---|
| | | Mean | S.D. | Mean | S.D. |
| 7 | {Day,Time} | 82.49 | 0.10 | 4613 | 559 |
| 24 | {Day,Time,3015} | 85.06 | 0.13 | 5083 | 467 |
| 25 | {Day,Time,2614} | 84.47 | 0.14 | 5351 | 887 |
| 47 | {Day,Time,3015,2614} | 87.23 | 0.13 | 6877 | 357 |
| | **Total Model Building Time (µs)** | | | **21924** | |

**Table 6-8: Accuracy of model evaluation and model building time for road link 2718 when using 4 context attributes**

| CtxSet | Context Attributes Applied | Accuracy (%) | | Model Building Time (µs) | |
|---|---|---|---|---|---|
| | | Mean | S.D. | Mean | S.D. |
| 7 | {Day,Time} | 89.99 | 0.07 | 4453 | 482 |
| 24 | {Day,Time,2515} | 98.85 | 0.02 | 4747 | 815 |
| 25 | {Day,Time,1401} | 90.22 | 0.09 | 4895 | 1821 |
| 47 | {Day,Time,2515,1401} | 98.97 | 0.01 | 4895 | 550 |
| | **Total Model Building Time (µs)** | | | **18992** | |

**Table 6-9: Model building time comparison between all context attributes and 4 context attributes applied**

| Road Segment ID to be Inferred | Model Building Time (µs) | | Reduction (%) |
|---|---|---|---|
| | All context attributes | 4 context attributes | |
| 1206 | 304678 | 19682 | 93.54% |
| 2613 | 318136 | 21924 | 93.11% |
| 2718 | 299445 | 18992 | 93.66% |

It can be seen that when applying only four context attributes (*day*, *time* and *traffic congestion degree of connected road link(s)*), the number of inference models is reduced to only four models and the model building time is reduced dramatically (presented in Table 6-9). The model building time reduction is over 93% while the accuracy remains considerably high as can be seen in Table 6-6, Table 6-7 and Table 6-8.

Our final reduced set of influential context attribute (*RS*) is thus composed of only *day*, *time* and the *traffic congestion degree of connected road link(s)*. This reduced set of influential context attributes (*RS*) is used to produce our refined CATE framework, explained in the next section.

## 6.5 Refinement of the CATE Framework

The design of the refined CATE framework presented in this chapter is our final artefact and is an extension and adjustment of the initial framework described in Chapter 4. In this section, we highlight only the parts that are improved from the initial CATE framework.

The process of refining the CATE framework began by collecting the influential context attributes and analysing these to obtain the reduced set of influential context attributes that we discussed in Section 0. We then analysed these and created the final reduced set of influential context attributes (*RS*) suitable for Bangkok. However, for other cities with different environment and contexts, the reduced set of influential context attributes may be different. Nonetheless, the process that we used to obtain the *RS* for Bangkok can also be applied to determine the *RS* relevant to other cities.

The refined CATE framework still consists of two phases, the *inference model building phase* and the *real time traffic inference phase*. This is the same as the initial framework, illustrated in Figure 4-8 in Chapter 4. The influential context attributes used in the initial framework were based on a review of existing research, recommendations from ITS researchers expert in ITS and the availability of data. However, in the refined CATE framework, the input context attributes for both phases have been narrowed down to members of the reduced set of influential context

attributes (*RS*) discussed in the previous section. *RS* is defined in the equation as follow.

$$RS = \{ \text{(day and time), traffic congestion degree of connected road links)} \}$$

The context attributes in *RS* and their domains of values are shown in Table 6-10.

**Table 6-10: Context Attribues in *RS* and their Domains of Value**

| Context Attributes | Domain of Value | Description |
|---|---|---|
| Traffic congestion | L, M, H | L = low congestion<br>M = medium congestion<br>H = high congestion |
| Day of the week | D1, D2, …, D7 | D1 = Sun, D2 = Mon, D3 = Tue, D4 = Wed, D5 = Thurs, D6 = Fri, D7 = Sat |
| Time of day | p1, p2,…, p24 | Time of the day is allotted into 24 period (p1 – p24), 1 hour per 1 period. |

The process for the refined framework starts by building $IM_{RS}$ inference models for the different sets of context attributes generated by combinations of context attributes in *RS* in the learning phase. Each model is created for each set of real time acquirable context attributes. A suitable inference model will be chosen in real time whenever the sensory data is detected as missing. The inference model selection depends on the context attribute(s) that can be acquired at that particular time. Once a suitable inference model is selected, the real time acquirable context attributes are used as inputs for the selected inference model to generate an *inferred traffic congestion degree*. The *inferred traffic congestion degree* is then used to compensate for the missing sensory data in the traffic dissemination system. In summary, the concept remains the same as that of the initial framework, but the set of influential context attributes has been pared down to *RS*.

The overall process of traffic congestion degree estimation is the same as shown in Figure 4-9, but the number of inference models is reduced from $IM_M$ to $IM_{RS}$. The

selected machine learning algorithm is J48, whose suitability for our framework was discussed in Chapter 4.

### 6.5.1 Inference Model Building of the Refined CATE Framework

The inference model building phase is a learning stage to create a suitable inference model for each set of context attributes. One model matches one set of run-time acquirable context attributes similar to the initial CATE framework. However, the number of influential context attributes applied and the number of models built is fewer than that of the initial framework.

Figure 6-8 describes the learning stage in the final CATE framework. The historical contexts are passed through the context attribute extractor and extracted to create the influential context attributes that are members of *RS*.

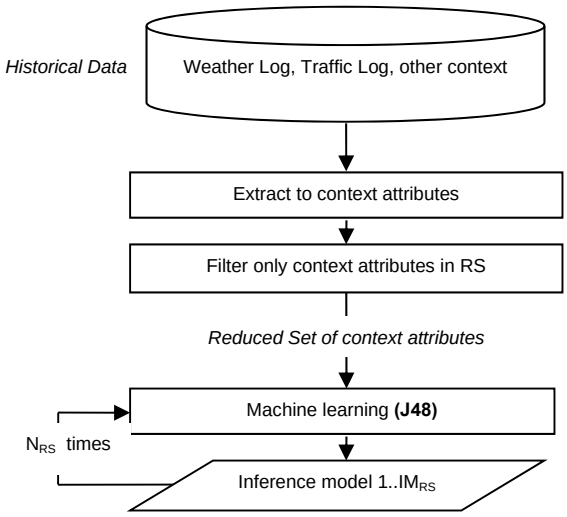



**Figure 6-8: Flow of the Learning Phase in the Refined Framework**

The filtered influential context attributes become inputs for the machine learning module to generate appropriate inference models. The total number of inference models within the refined framework is now equal to $N_{RS}$ as defined in Equation eq.1

below. The model building process is repeated $N_{RS}$ times for all context attributes in *RS*.

$$\mathrm{N}_{RS} = \sum_{k=1}^{n}(C_k^n - n(A)),\ \text{for all}\ \mathrm{N}_{RS} \in I^+ \qquad \text{(eq.1)}$$

In Equation eq.1, $N_{RS}$ is the total number of inference models, $C_k^n$ is the combination of context attributes in *RS*, $n(A)$ is the number of $\{\mu'_{(day,time)}\}$, μ is the union of combinations of context attributes in *RS* that are composed of both *day* and *time* context attributes, *n* is the total number of influential context attributes in *RS*, and *k* is the number of member(s) in each context attribute set and 1<= *k* <= *n*.

In Chapter 4, we determined that J48, which is the C4.5 decision tree machine learning algorithm, is the algorithm best suited to our proposed framework. The J48 machine learning algorithm is thus used for building the models. Figure 6-8 depicts the flow of the learning phase for the refined framework.

The inputs for the model building are the historical contexts that are also members of *RS*. Members of *RS* can consist of the sensory traffic data of connected roads and the day and time obtained from the time-stamp of the traffic log. It is noticeable that, with the refined design, the cost (in terms of computing resources and hardware required, effort, human resources and time) and the difficulty of acquiring the necessary context attributes for the learning process are reduced. Similar to the initial framework, the final artefact is designed to be self-learning. The machine learning task of building inference models can be repeated at user-defined time intervals to update the inference models to obtain dependable inference results.

### 6.5.2 Inference Phase of the Refined CATE Framework

The process of real time traffic inference of the final framework is the same as that of the proposed framework in Figure 4-12 in Chapter 4, but the context attributes used for inferring the missing sensory traffic data are now only those in the *RS*. Similar to the initial CATE framework, each road segment has its own intelligent RoadLink Module as shown in Figure 4-13. Each RoadLink module can communicate with a

RoadLink module of connected road segments, so they can exchange sensory traffic data with each other.

The refined CATE framework has the same mechanics as the initial CATE framework. When sensory traffic data is available, the CATE framework reports actual traffic conditions based on real time data from traffic sensors. If the traffic data of a particular road segment is lost, it then adaptively starts the inference phase to infer the traffic congestion degree based on discoverable external contexts using selected pre-created inference models. The appropriate inference models obtained from the learning stage are used for calculating the inferred traffic congestion degrees to compensate for missing sensory data and to report the congestion level to road users. At particular times, the CATE framework attempts to collect the current available context attributes that are in the *RS*, as much as it can, to use as inputs to the inference process. The collected contexts are then converted to context attributes before being used as inputs for a suitable inference model. The inference model is selected from the total $N_{RS}$ inference models based on the context attributes acquired at real time. The inferred traffic congestion degree is then generated by the selected inference model.

The CATE framework is set to automatically relearn at predefined time intervals (for example, every month) as explained in section 4.4.3 of Chapter 4. As time passes, the framework keeps new incoming actual data along with other contexts in *RS* for future relearning purposes. The inferred values are not included in the relearning task in order to avoid skewed results in the future.

It can be seen that the algorithm of the inference process remains the same as that described in Figure 4-11. The difference lies in the reduced set of context attributes used – context attributes must be members of *RS* – and thus the number of inference models has been reduced from $N_m$ to $N_{RS}$.

## *6.6 An Evaluation of the Refined CATE Framework (Evaluation II)*

For the evaluation in this chapter, we assess the improvement of the refined CATE framework over the initial CATE framework. We evaluate the model building time

and framework costs. Costs include the required hardware, computing resources, human resources, effort and time. Implementing a good quality system while reducing its costs can contribute to the success of the information technology system [147, 148]. In addition, we also develop a program that realizes the refined proposed framework to ensure that the proposed framework is feasible and efficient. We also evaluate the performance of the accuracy of the refined CATE framework compared to the initial CATE framework and two other approaches (the mode and single model approaches).

We implement the program using the Groovy programming language [127] and use Weka classes within the program similar to the implementation in Chapter 5. However, in this implementation, we use the context attributes from the reduced set of influential context attributes (*RS*). The details of the programming code can be found in Appendix B.

The source of data and the data preparation for Evaluation II are the same as those explained in Section 5.1 of Chapter 5, but in this chapter we select only the context attributes that appear in the reduced set of influential context attributes (*RS*) that we analysed and identified in Section 0 of this chapter.

The experiment setup and the method of simulating a missing data situation for our test data set are the same as explained in Section 5.2.1, but again we select only context attributes from *RS* instead of using all the context attributes.

In this second evaluation, we also implement three different approaches similar to Evaluation I in Chapter 5. Those approaches are the mode approach (see Section 4.3.1), the single model approach (see Section 4.3.2) and the refined CATE framework. The implementation of the mode approach is the same as that explained in Section 5.3.1 The Mode Approach. We then compare the result.

The results from our evaluation show that the accuracy and inferring time of the mode approach are the same as those reported in Chapter 5 in Section 5.4.1. The results are the same because reducing the number of context attributes has no effect under this approach. When sensory traffic data is missing, the framework always replaces the missing data with a mode value in order to report the inferred traffic congestion

degree. The comparison of using the mode approach with our refined CATE framework can be found in Section 6.6.2.

The implementation of the single model approach is also the same as that explained in Section 5.3.2 The Single Model Approach, but the context attributes used to build an inference model and infer missing data are context attributes drawn only from *RS*. In this chapter, we show only the comparison of the performance of the single model approach (when using *RS*) and our refined CATE framework in Section 6.6.2. The results of our experiment from the single model approach when using only context attributes from *RS* can be found in Appendix A.

The implementation of the refined CATE framework is the same as that explained in 5.3.3 The CATE Framework Approach, but the context attributes used for learning and for building inference models are reduced to only the context attributes from *RS*. In addition, the number of inference models to be built is reduced to $N_{RS}$ (see equation 1 in section 0). The context attributes used for inferring the missing sensory traffic data are also only members of *RS*.

### 6.6.1 Results from the Refined CATE Framework

In this section, we report the results from implementing our refined CATE framework. We also evaluate the improvement of the refined CATE framework over the initial CATE framework. Furthermore, we compare the performance of the refined CATE framework with the performance of other approaches (the mode approach and the single model approach) in this section.

The process starts with building inference models. The built models are then evaluated using a stratified 10-fold cross validation method. We build and evaluate models ten times (the accuracy is different each time) and calculate the average and standard deviation (SD) of accuracy as well as the average model building time.

Table 6-11, Table 6-12, and Table 6-13 show the averages for accuracy and the time taken to build the models with the refined CATE framework for different road segments. The number of models built is reduced from the implementation of the initial CATE framework in Chapter 5. The model building time is thus decreased dramatically from the implementation of the initial CATE framework in Chapter 5

because the number of context attributes applied is reduced. The accuracy and the model building time shown in Table 6-11, Table 6-12, and Table 6-13 is slightly different from the results reported in Table 5-15, Table 5-16, and Table 5-17 when considering the same context attribute set because we ran new experiments after RS was selected. Table 6-14 summarises the comparison between the initial CATE framework and the refined CATE framework in terms of the time taken to build the models. This comparison is also presented graphically in Figure 6-9.

**Table 6-11: Average of accuracy and model building time of model evaluation for refined CATE approach for road link 1206 when applying context attributes in *RS***

| CtxSet | Context Attributes Applied | Accuracy (%) | | Models Building Time (µs) | |
|---|---|---|---|---|---|
| | | Mean | S.D. | Mean | S.D. |
| 1 | {Day,Time} | 90.61 | 0.11 | 6216 | 983 |
| 2 | {Day,Time,1211} | 92.70 | 0.06 | 6764 | 914 |
| 3 | {Day,Time,1403} | 92.08 | 0.12 | 6344 | 295 |
| 4 | {Day,Time,1211,1403} | 93.75 | 0.08 | 6837 | 794 |
| | **Total Model Building Time (µs)** | | | **26160** | |

**Table 6-12: Average of accuracy and model building time of model evaluation for refined CATE appraoch for road link 2613 when applying context attributes in *RS***

| CtxSet | Context Attributes Applied | Accuracy (%) | | Models Building Time (µs) | |
|---|---|---|---|---|---|
| | | Mean | S.D. | Mean | S.D. |
| 1 | {Day,Time} | 82.49 | 0.10 | 5428 | 845 |
| 2 | {Day,Time,3015} | 85.06 | 0.13 | 7176 | 1523 |
| 3 | {Day,Time,2614} | 84.47 | 0.14 | 9198 | 1244 |
| 4 | {Day,Time,3015,2614} | 87.23 | 0.13 | 8327 | 1265 |
| | **Total Model Building Time (µs)** | | | **30129** | |

**Table 6-13: Average of accuracy and model building time of model evaluation for refined CATE appraoch for road link 2718 when applying context attributes in *RS***

| CtxSet | Context Attributes Applied | Accuracy (%) | | Models BuidlingTime (µs) | |
|---|---|---|---|---|---|
| | | Mean | S.D. | Mean | S.D. |
| 1 | {Day,Time} | 89.99 | 0.07 | 5166 | 795 |
| 2 | {Day,Time,2515} | 98.85 | 0.02 | 6685 | 998 |
| 3 | {Day,Time,1401} | 90.22 | 0.09 | 6867 | 1344 |
| 4 | {Day,Time,2515,1401} | 98.97 | 0.01 | 6658 | 1611 |
| | **Total Model Building Time (µs)** | | | **25376** | |

**Table 6-14: Model building time comparison between initail CATE framework and refined CATE framework**

| Road Segment ID to be Inferred | Model Building Time (μs) | | Reduction (%) |
|---|---|---|---|
| | All context attributes | RS | |
| 1206 | 304678 | 26160 | 91.41% |
| 2613 | 318136 | 30129 | 90.53% |
| 2718 | 299445 | 25376 | 91.53% |

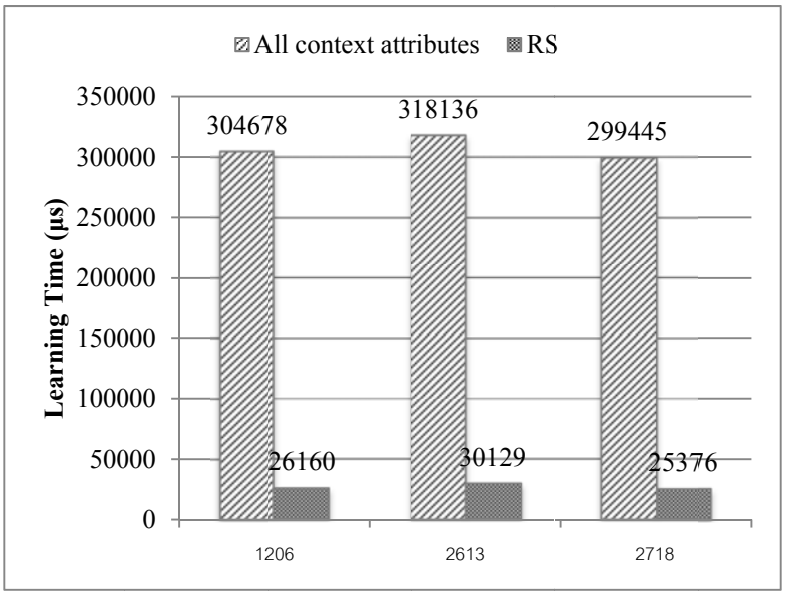



**Figure 6-9: Model building time comparison between the initial CATE framework and the refined CATE framework**

After building $N_{RS}$ inference models for each context attribute set, these inference models are used to infer the missing sensory traffic data of the observed road segments. The real situation is simulated as described in Section 5.2.1 (but the context attributes used are now the context attributes that appear in *RS*).

Similar to the single model approach, we set the missing data rate to 20%, 40%, 60%, 80% and 100% and present the consequent results in this section. We also run the experiment ten times and present the average accuracy and average inferring times (the length of time taken to produce the inferred traffic congestion degree) for the ten runs in this section.

Table 6-15 and the graph in Figure 6-10 illustrate the average time used to infer the traffic congestion degree. Table 6-16 and the graph in Figure 6-11 show the average accuracy and SD at different missing data rates for each road segment. In addition, similar to the mode approach and the single model approach, we also present the results for active and non-active periods respectively in Table 6-17 and Table 6-18 and illustrate them in Figure 6-12 and Figure 6-13.

**Table 6-15: Inferring time (μs) for specific missing data rates (%) for the refined CATE framework approach when applying *RS***

| Road Link | Inferring Time (μs) at specific data missing rate (%) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 20% | | 40% | | 60% | | 80% | | 100% | |
| | Mean | S.D. | Mean | S.D. | Mean | S.D. | Mean | S.D. | Mean | S.D. |
| 1206 | 1412 | 30 | 1282 | 9 | 1180 | 7 | 1086 | 6 | 1001 | 4 |
| 2613 | 1927 | 36 | 1654 | 16 | 1414 | 16 | 1251 | 5 | 1132 | 10 |
| 2718 | 1197 | 16 | 1171 | 8 | 1140 | 7 | 1100 | 15 | 1043 | 7 |

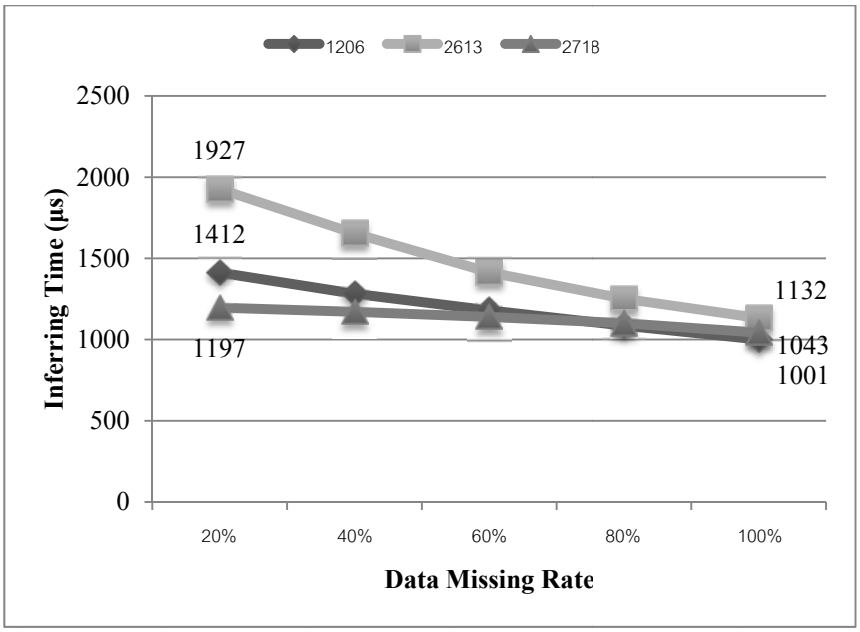



**Figure 6-10: Inferring time (μs) for specific missing data rates (%) for the refined CATE framework approach when applying *RS***

**Table 6-16: Accuracy (%) at specific missing data rates (%) for the refined CATE framework approach when applying *RS***

| Road Link | Accuracy (%) at specific missing data rate (%) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 20% | | 40% | | 60% | | 80% | | 100% | |
| | Mean | S.D. | Mean | S.D. | Mean | S.D. | Mean | S.D. | Mean | S.D. |
| 1206 | 88.72 | 0.23 | 88.90 | 0.29 | 88.78 | 0.14 | 88.49 | 0.08 | 88.43 | 0.00 |
| 2613 | 81.96 | 0.22 | 81.75 | 0.30 | 81.64 | 0.18 | 81.27 | 0.21 | 80.04 | 0.00 |
| 2718 | 96.73 | 0.17 | 94.96 | 0.18 | 93.15 | 0.17 | 91.61 | 0.17 | 89.47 | 0.00 |

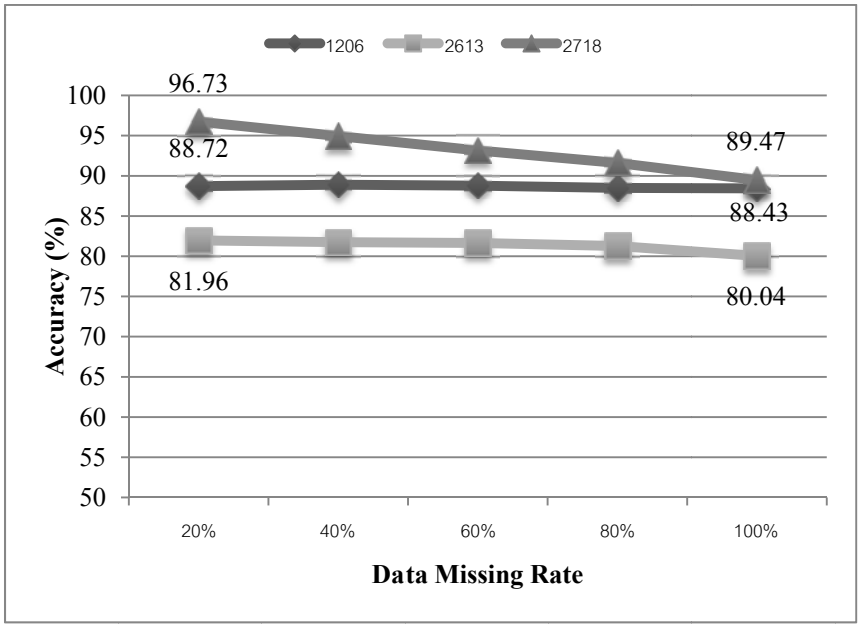



**Figure 6-11: Accuracy (%) at specific missing data rates (%) for the refined CATE framework approach when applying *RS***

**Table 6-17: Accuracy (%) in *active period* at specific missing data rates (%) for the refined CATE framework approach when applying *RS***

| Road Link | Active period accuracy (%) at specific missing data rates (%) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 20% | | 40% | | 60% | | 80% | | 100% | |
| | Mean | S.D. | Mean | S.D. | Mean | S.D. | Mean | S.D. | Mean | S.D. |
| 1206 | 81.31 | 0.47 | 81.56 | 0.35 | 81.41 | 0.24 | 80.95 | 0.18 | 80.62 | 0.00 |
| 2613 | 65.14 | 0.47 | 64.55 | 0.63 | 64.15 | 0.26 | 63.41 | 0.43 | 61.65 | 0.00 |
| 2718 | 93.95 | 0.29 | 90.58 | 0.39 | 87.21 | 0.30 | 84.23 | 0.35 | 80.06 | 0.00 |

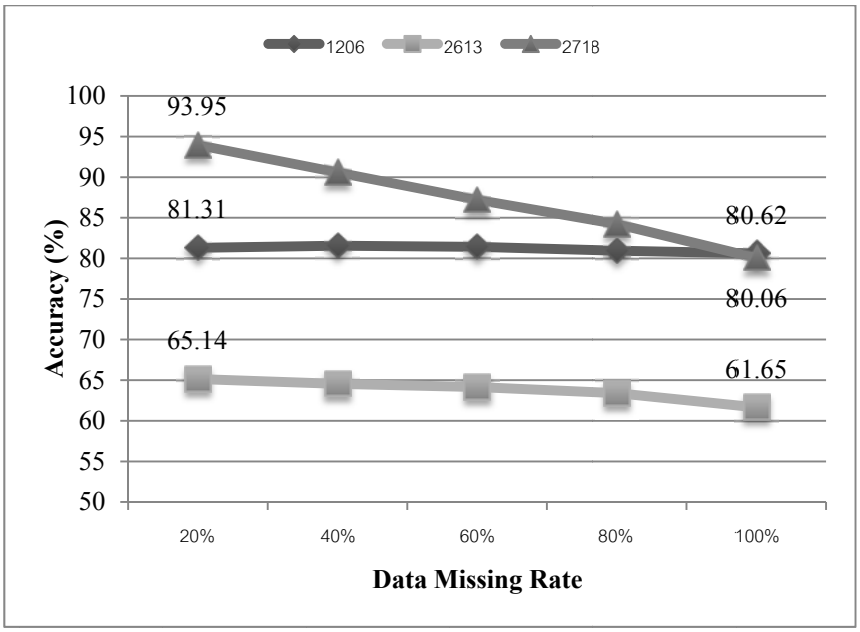



**Figure 6-12: Accuracy (%) in *active period* at specific missing data rates (%) for the refined CATE framework approach when applying *RS***

**Table 6-18: Accuracy (%) in *non-active period* at specific missing data rates (%) for the refined CATE framework approach when applying *RS***

| Road Link | Non-active period accuracy (%) at specific missing data rates (%) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 20% | | 40% | | 60% | | 80% | | 100% | |
| | Mean | S.D. | Mean | S.D. | Mean | S.D. | Mean | S.D. | Mean | S.D. |
| 1206 | 95.97 | 0.12 | 96.09 | 0.20 | 95.99 | 0.11 | 95.90 | 0.10 | 96.11 | 0.00 |
| 2613 | 98.58 | 0.16 | 98.70 | 0.11 | 98.82 | 0.07 | 98.84 | 0.07 | 98.12 | 0.00 |
| 2718 | 99.50 | 0.11 | 99.27 | 0.10 | 99.00 | 0.06 | 98.87 | 0.04 | 98.73 | 0.00 |

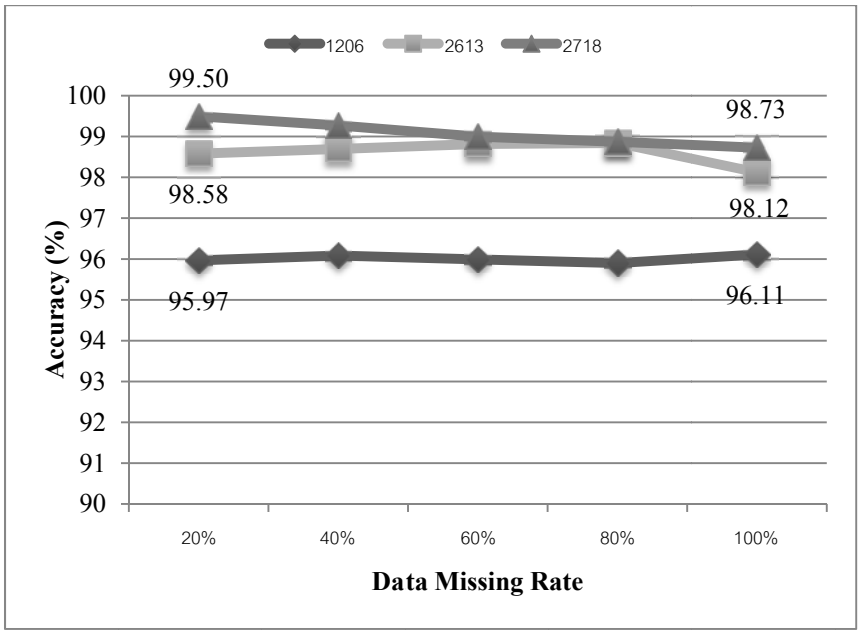



**Figure 6-13: Accuracy (%) in *non-active period* at specific missing data rates (%) for the refined CATE framework approach when applying *RS***

## 6.6.2 Discussion

In this section, we discuss the results from the refined CATE framework against the results from the initial CATE framework. Since the number of inference models to be built in the refined CATE framework is dramatically reduced from the sixty-three models in the initial CATE framework (see Table 5-16, Table 5-15 and Table 5-17) to only four models in the refined CATE framework, the time required to learn and build the inference models is also dramatically reduced.

Table 6-14 and Figure 6-9 explicitly show the improvement of the refined CATE framework compared to the initial CATE framework. The time to learn and then build all inference models is greatly reduced. The model building time reduction is over 90%. The time required to build the models is reduced dramatically while the accuracy remains high even though the number of influential context attributes applied is reduced.

In addition to improving the model building time, compared to the initial CATE framework, the refined CATE framework can also improve over the initial CATE framework in terms of cost. These costs relate to the hardware, human resources, effort and time required to acquire the influential context attributes of the refined CATE framework. The requirement to install additional hardware such as rain sensors is also eliminated when the refined framework is employed. In addition, the difficulty and effort in communicating with other organizations to obtain rain data can also be eliminated.

The graph of inferring time for each road segment shown in Figure 6-10 (Inferring time (μs) for specific missing data rates (%) for the refined CATE framework approach when applying RS) and the graph in Figure 5-25 (Inferring times of the CATE framework approach at specific missing data rates) are similar shapes. However, the overall inferring time resulting from the refined CATE framework is reduced from the initial CATE framework.

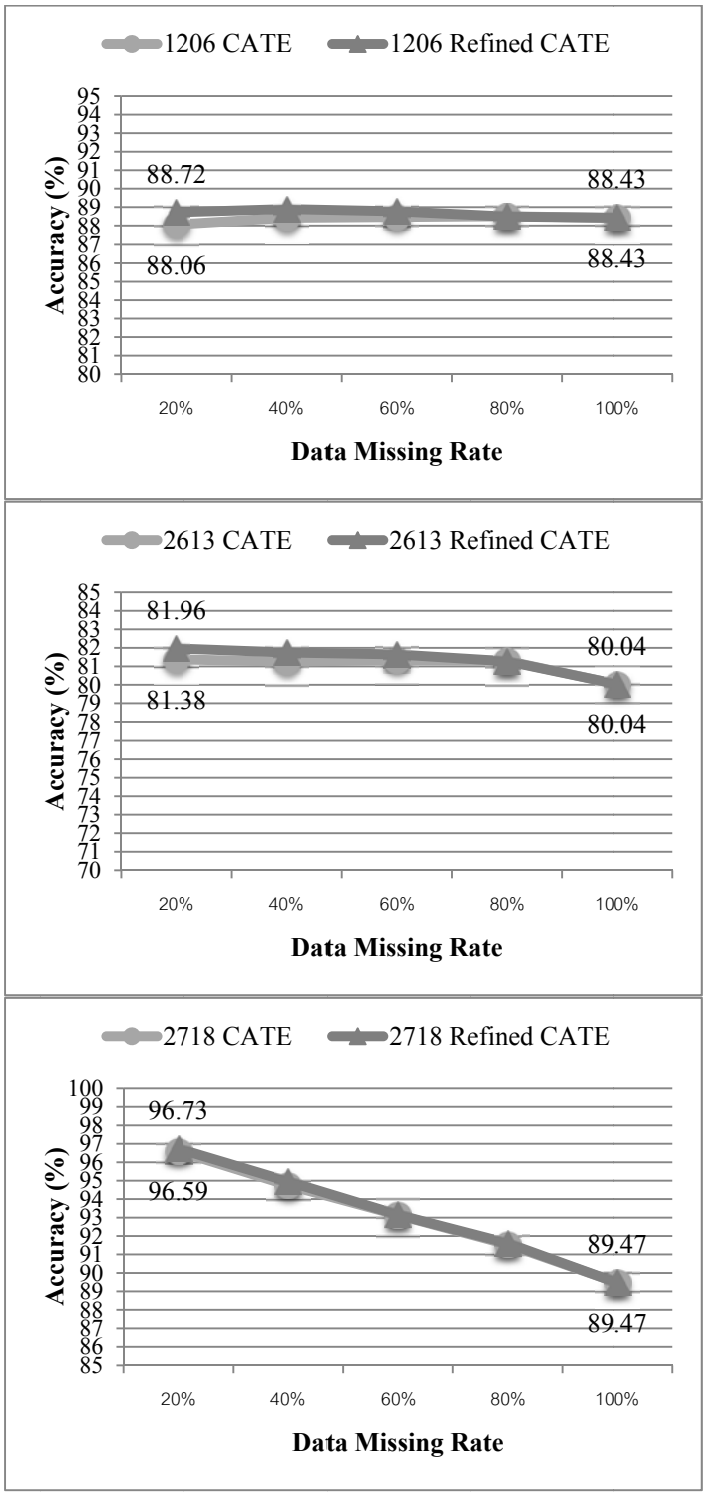



**Figure 6-14: Comparison of the initial CATE and refined CATE frameworks for accuracy at specific missing data rates (%) of road links 1206, 2613 and**

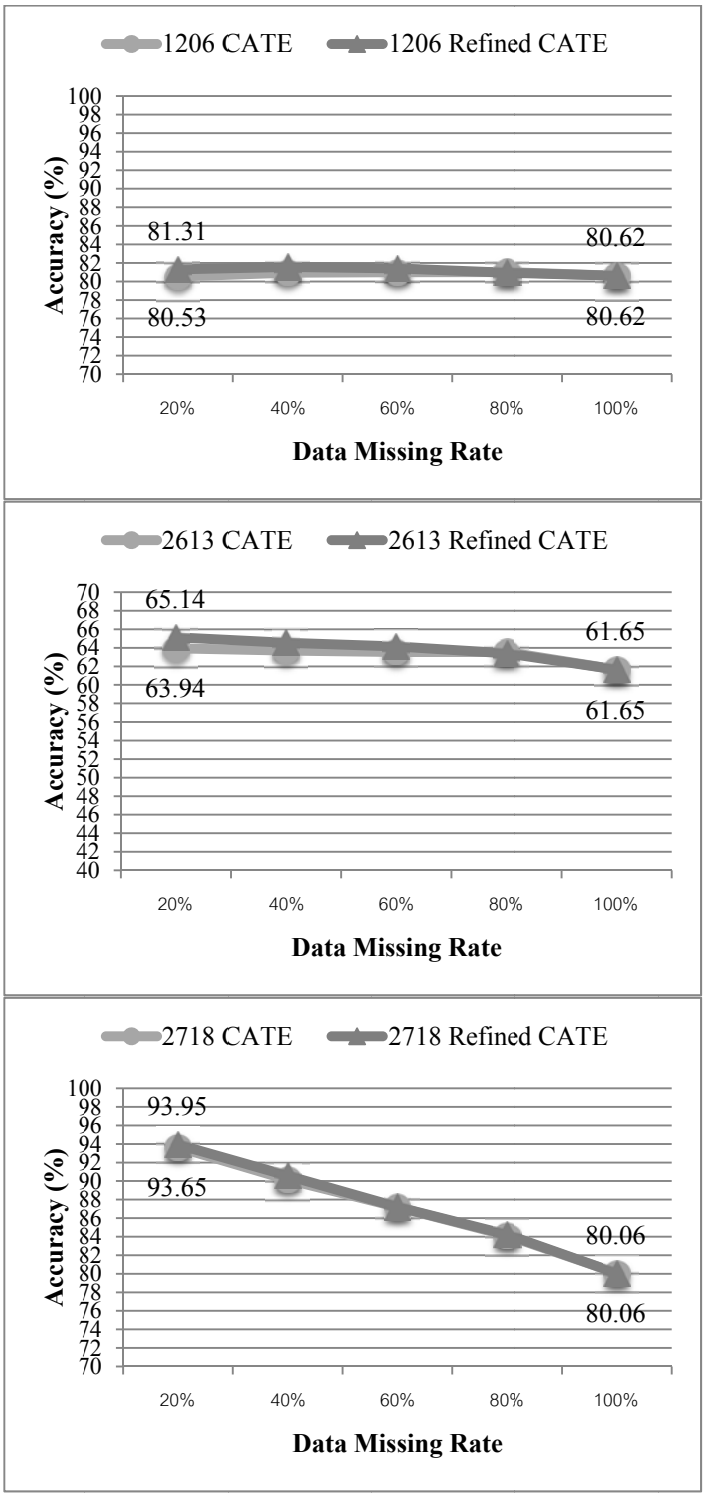



**Figure 6-15: Comparison of the CATE and refined CATE frameworks (*active period*) for accuracy at specific missing data rates (%) of road links 1206, 2613 and 2718**

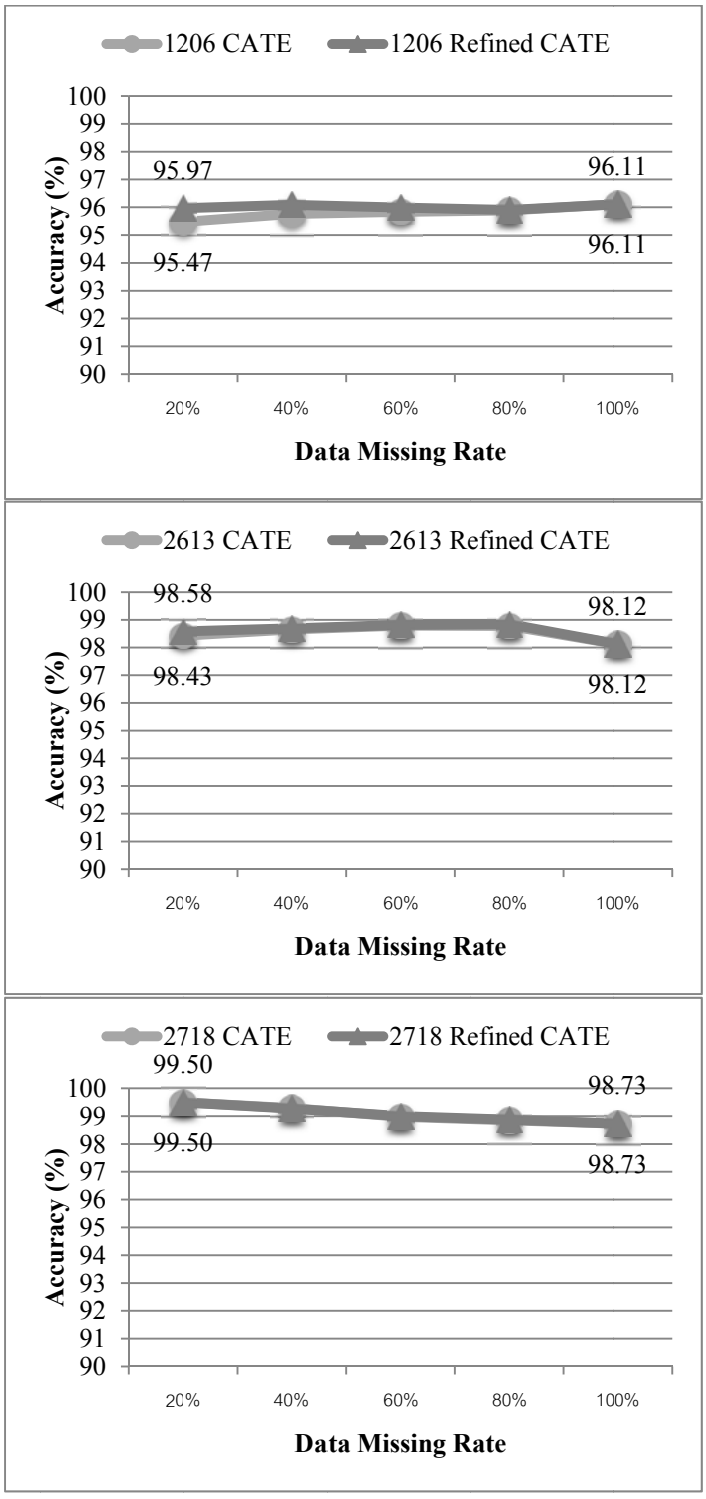



**Figure 6-16: Comparison of the CATE and refined CATE frameworks (*non-active period*) for accuracy at specific missing data rates (%) of road links 1206, 2613, and 2718**

When we compare the accuracy of the initial CATE and refined CATE framework as shown in Figure 6-14, Figure 6-15 and Figure 6-16, the accuracy of the refined CATE framework is the same or higher as the initial CATE framework even though the number of context attributes applied in the refined CATE framework is reduced. In addition, we notice that when the missing data rate is 100%, the accuracy is equal in both the initial and refined CATE frameworks. This is because when the missing data rate is 100%, it means only the inference model for *"day"* and *"time"* will be applied to infer the missing sensory traffic data and this inference model is the same for both the initial and refined CATE frameworks.

Our evaluation also shows that applying a greater number of influential context attributes does not necessarily improve accuracy. In contrast, it may result in unnecessary cost increases. As presented in [95] (p.20), most of the effort in the data mining process goes to data preparation. Using our reduced set of influential context attributes will reduce the effort of data preparation and effort in collecting data, thus reducing the overall cost of implementation since effort is considered a part of the cost [147]. Reducing the required context attributes also reduces the investment in extra hardware, reduces the effort to connect and coordinate the system across organizations and reduces the human resources required, all while maintaining high performance. Overall, we can conclude that the refined CATE framework is an improvement over the initial framework in terms of model building time and cost reduction.

In addition to the evaluation of the refined CATE framework against the initial CATE framework, we also evaluate our refined CATE framework against the mode approach and the single model approach. The following section shows the results from evaluating the inferring time and accuracy with different missing data rates in deployment when using each approach. Figure 6-17 compares the inferring time when using the single model approach and the refined CATE framework. The comparison of accuracy for the three approaches is shown in Figure 6-18. The graphs in Figure 6-19 and Figure 6-20 show the comparison of accuracy for the three approaches in both active and non-active periods respectively.

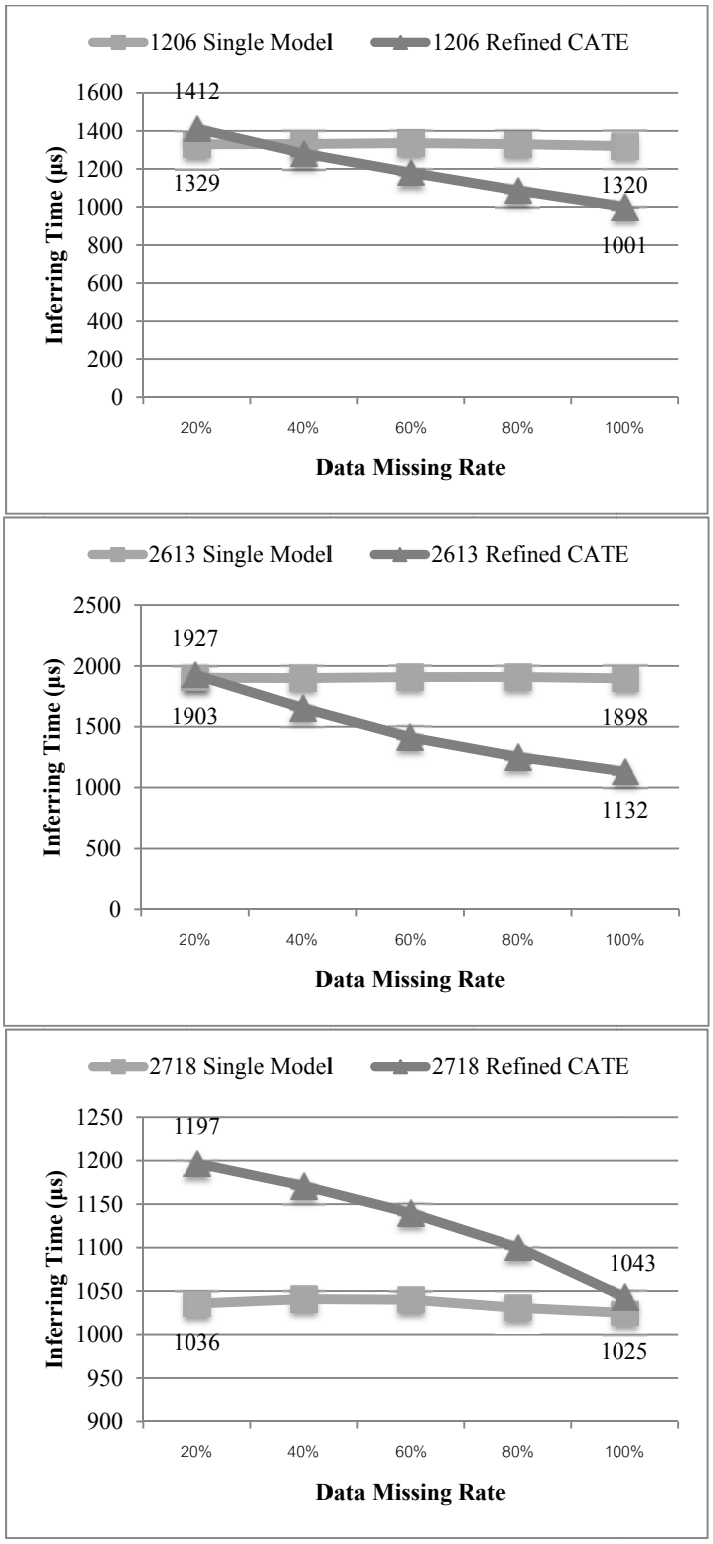



**Figure 6-17: Comparison of the single model approach and the refined CATE framework approach in inferring time (µs) at specific missing data rates (%) of road links 1206, 2613 and 2718**

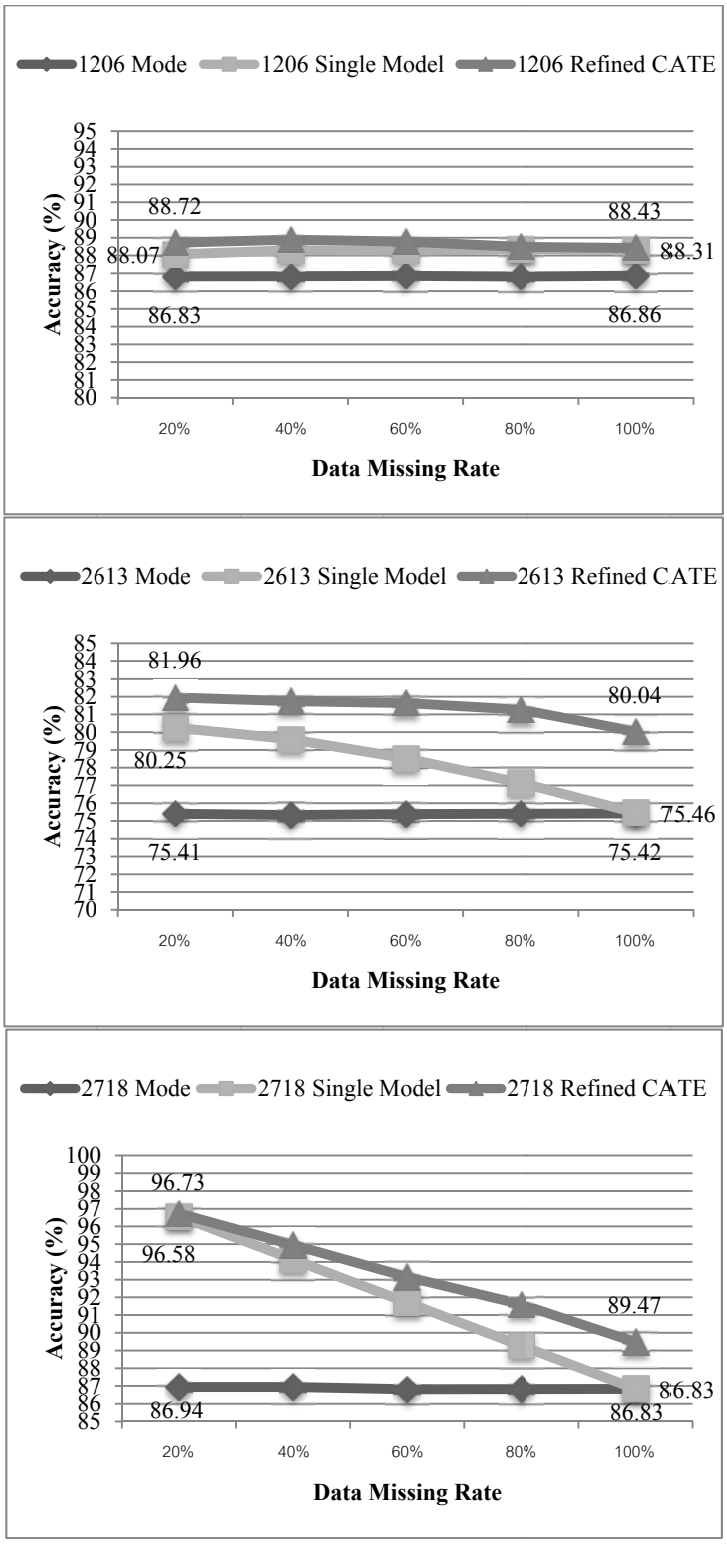



**Figure 6-18: Comparison of the mode approach, single model approach and refined CATE framework approach in accuracy (%) at specific missing data rates (%) of road links 1206, 2613 and 2718**

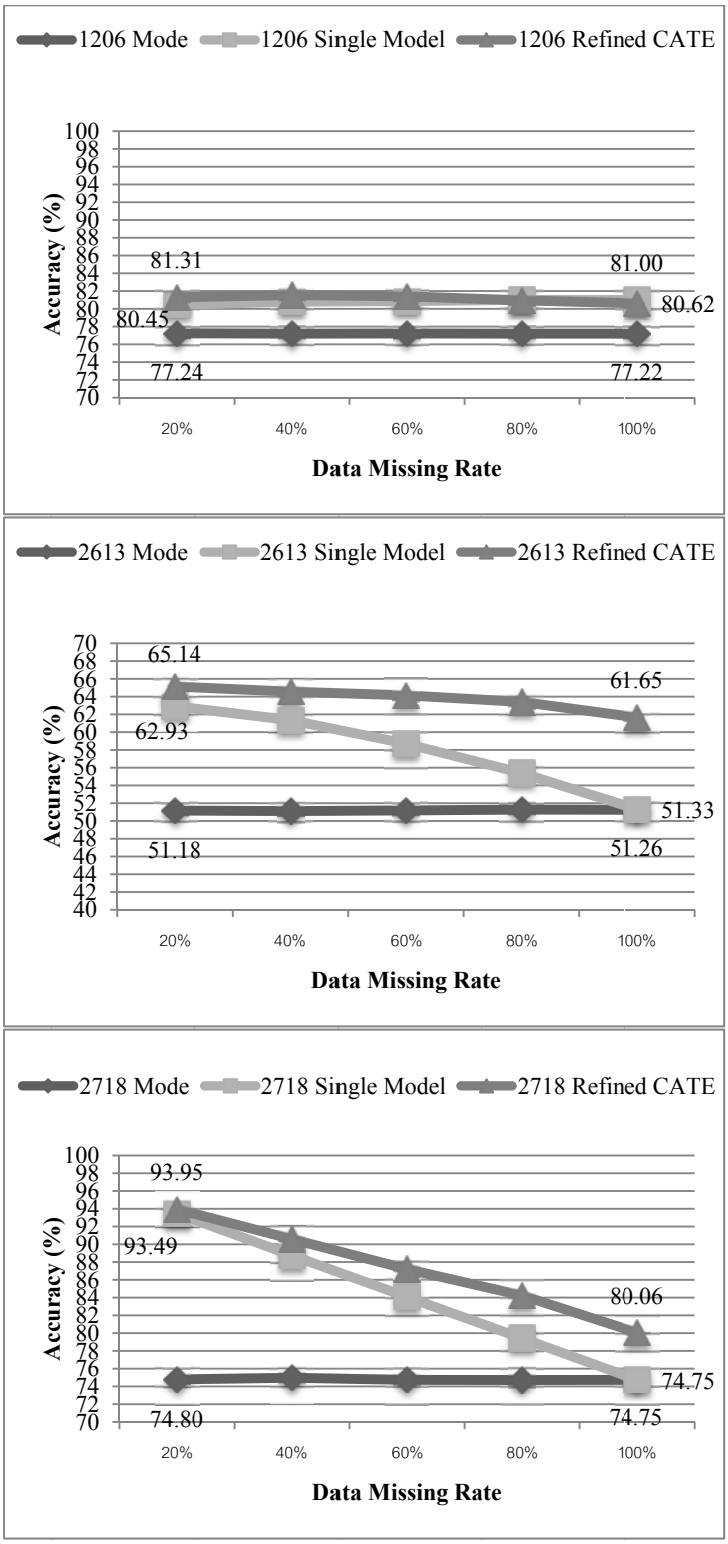



**Figure 6-19: Comparison of the mode approach, single model approach and refined CATE framework approach (*active period*) in accuracy (%) at specific missing data rates (%) of road links 1206, 2613 and 2718**

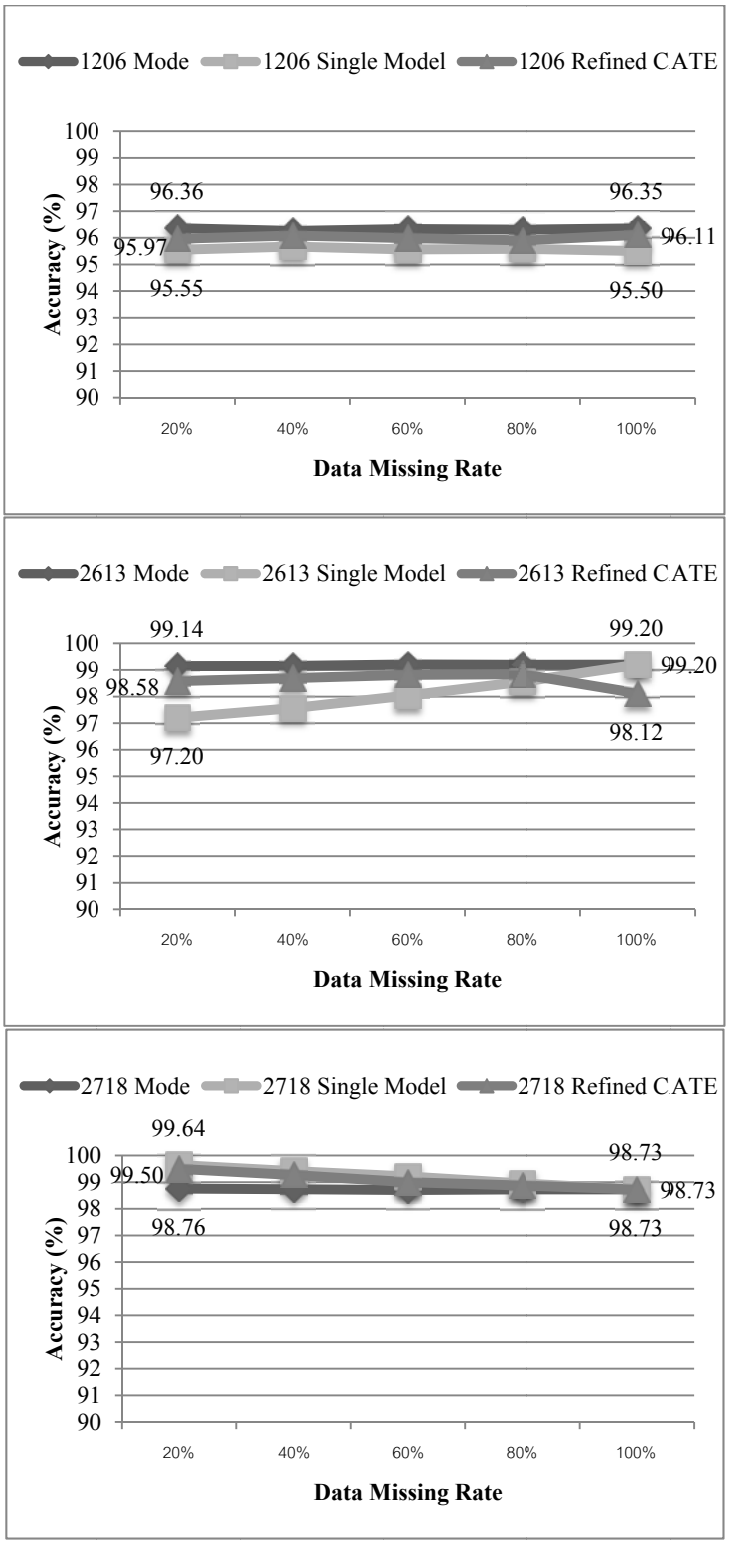



**Figure 6-20: Comparison of the mode approach, single model approach and refined CATE framework approach (*non-active period*) in accuracy (%) at specific missing data rates (%) of road links 1206, 2613, and 2718**

Figure 6-17 illustrates the comparison of the inferring times of the single model approach and our refined CATE framework. The inferring time of the mode approach is less than 0.2 μs because the calculation is very simple. When the missing sensory traffic data is detected, the system always reports the mode value. However, its lack of accuracy renders the mode approach unsuitable, especially in active periods, as we explained in Section 5.4.1.

The inferring time of the single model approach is unchanged for all missing data rates because the size of the inference model (the decision tree) is constant because the single model approach uses only one inference model for all cases. In contrast, the inferring time of the refined CATE framework decreases as the missing rate increases. It is noticeable that the shape of the graph shown in Figure 6-17 is the same as the shape of the graph shown in Figure 5-29. The inferring time of the refined CATE framework reduces as the missing data rate climbs, because as the number of missing context attributes increases, the system chooses models that are built for fewer context attributes, and thus the size of the model (the size of the decision tree) is smaller. In addition, the inferring time for the refined CATE framework is also less than that of the initial CATE framework.

Similar to Evaluation I in Chapter 5, the graphs in Figure 6-18 prove that our refined CATE framework still performs better than the mode approach and the single model approach for accuracy at different missing data rates. Our refined CATE framework gives higher accuracy than other approaches for all missing data rates. Even though the accuracy drops as the missing data rate increases, it is still over 80%. Our refined CATE framework is resilient to high missing data rates even at 100% (or when only the *day* and *time* context attributes are available). Furthermore, the accuracy (in deployment) of the refined CATE framework compared to the initial CATE framework does not decline even though the context attributes applied are reduced to *RS*.

Similar to the initial CATE framework, the superior performance of the refined CATE framework is obvious even when comparing the accuracy of the active period with the other approaches in Figure 6-19. The experiment results also prove that our CATE framework is resilient to high missing data rates even though no context attributes

other than *day* and *time* are acquirable (day and time are known fact that always available).

In addition, the refined CATE framework can be used even when collected historical data is sparse. Our evaluation shows that the refined CATE framework can give high accuracy even when the historical data for learning and building models includes only 10,000 records (about one month's worth of records). However, the framework becomes smarter when new actual data is collected because the framework incorporates a relearning process and rebuilds models at pre-defined time periods. In our experiment relearning is set at every 10,000 records, or about once a month. The optimal time period for relearning is left for future research. Thus, our refined framework can also work well with road segments with limited historical data such as road segments with newly installed traffic sensors.

Overall, when comparing the performance of the refined CATE framework with other approaches (the mode approach and the single model approach), its accuracy in deployment is superior while the processing time of the refined CATE system is reduced. This demonstrates the efficiency of our refined CATE framework and its improvement over the initial CATE framework.

## 6.7 Chapter Summary

This chapter presents the refinement of our proposed framework and its evaluation. The chapter begins with an analysis of survey data to confirm the utility of traffic information for Bangkok road users (which also confirms the usefulness of our proposed framework) and to confirm our selection of influential context attributes. This is followed by the analysis of the reduced set (*RS*), which is the set of context attributes selected for our refined CATE framework. Our reduced set analysis method is also useful in helping traffic management organizations decide the influential context attributes that warrant investment. Such process can also be applied to determine the *RS* relevant to other cities. We then illustrate the design of the framework and follow with the evaluation.

Evaluation II in this chapter assesses the final artefact (the refined CATE framework) mainly on its accuracy and processing time. We include a discussion on the resources

and costs required to implement the framework. The results from the evaluation demonstrate that the refined framework is an improvement over the initial framework. The processing time is decreased significantly while the accuracy remains high. The costs of implementing the improved framework (including hardware and human resources, effort and time) are also reduced. Hence, the overall performance of the framework is improved. Moreover, we also show that a system implementing our improved framework can cope well with missing context attributes even when the sensor failure rate is up to 100%. The refined CATE framework also performs better than the mode and the single model approaches in terms of accuracy. Overall the refined CATE framework, which is our final artefact, accomplishes good performance with reduced costs.

Evaluation II in this chapter proves that our final artefact can answer our research questions. It performs well in addressing the problem of missing sensory traffic data. In addition, Evaluation II shows the improvement of the final artefact from the initial framework and its superior performance compared to other approaches.

The next chapter describes the next stage of our research: a study into the potential use of social networks for TIS and the traffic information needs of Bangkok's road users. The results from this study have relevance both to improving the implementation of TIS and traffic information dissemination in the future.

# Chapter 7 Traffic Information Usage of Bangkok's Road Users and the Potential Use of Social Networks for Traffic Information Systems in Bangkok

In addition to the report on the factors that influence to Bangkok's road traffic in the perception of Bangkok's road users (reported in Chapter 6), in this chapter we report the traffic information needs of these users and their preferred communication channels for this information. We also report the potential use of social networks to yield traffic data and for traffic information dissemination. The statistical analysis and discussion of survey outcomes presented in this chapter result in suggested guidelines to help the design of future efficient TIS and traffic report services. The survey methodology was described in Chapter 6.

## *7.1 Traffic Information Usage*

In this section, we present the survey results and the analysis of the traffic information needs of Bangkok's road users in terms of the type and source of traffic information required, and how this information is used in route planning.

### 7.1.1 Primary Sources of Information

The results in this section (illustrated in Figure 7-1) show how Bangkok road users consume traffic data, and in particular, the media channels and sources accessed by road users. Most drivers usually receive traffic information from radio (20%). The electronic board or variable message sign (VMS), which is installed on major roads in Bangkok, is ranked second in terms of access (15.2%). Applications on mobile

devices are ranked third (12.7%). We expected this number to be higher as there has been a rapid growth in the number of smartphone users in Thailand [149, 150]. However, this low number may be due to the limited number of smartphone traffic information applications available when the survey was taken.

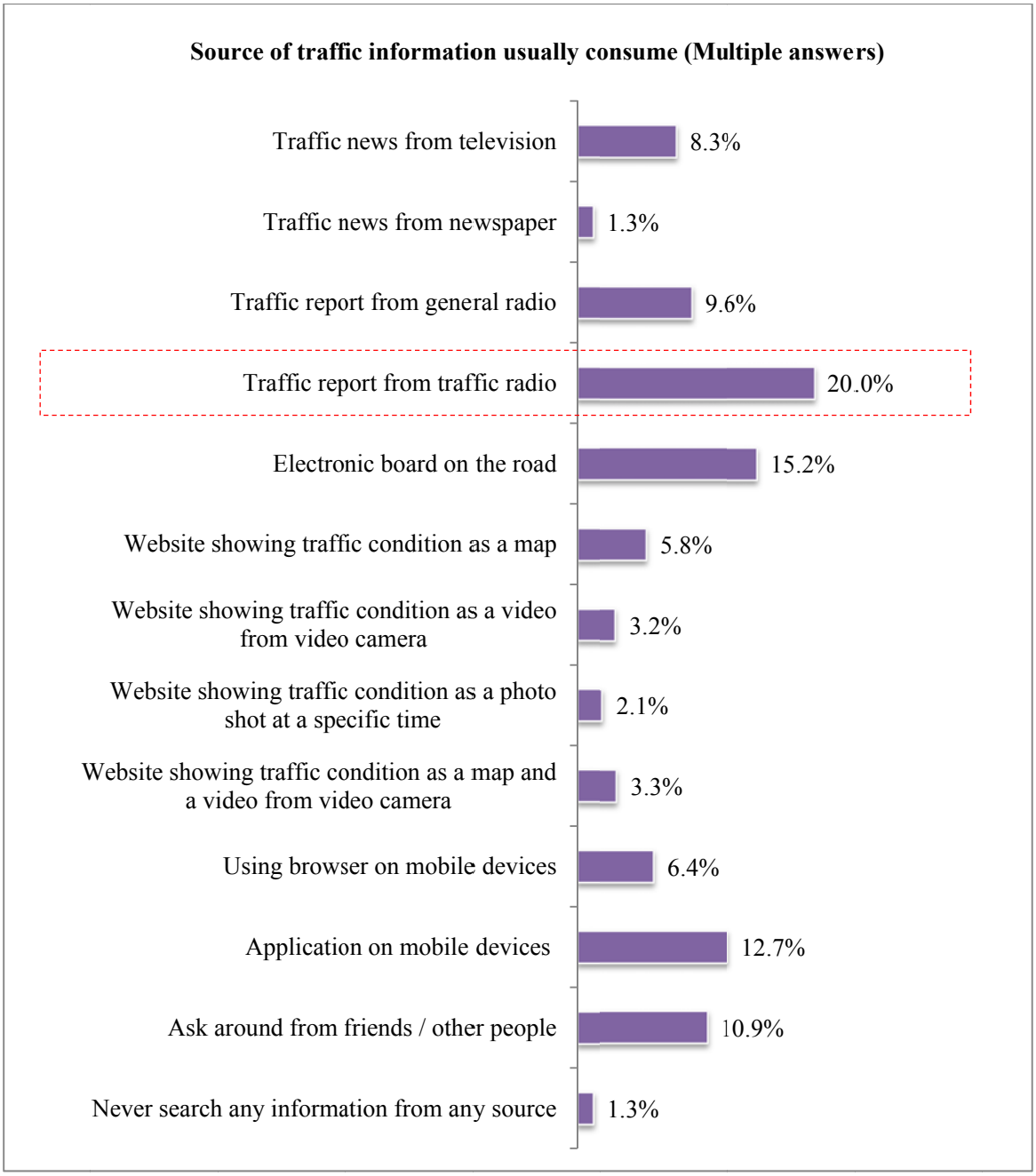



**Figure 7-1: Traffic information sources of road users**

When grouping results by media, it can be seen that most respondents consume traffic information via radio (29.6%). The next largest group belongs to those that use mobile devices (19.1%), which is the combination of both applications and browsers on mobile devices. In addition, 14.4% of respondents choose to access traffic information via websites.

### 7.1.2 Traffic Information Sought by Users

When we asked about the type of information desired by respondents, and at what point in their journey they would like to receive such information, we found that most respondents search for a route before starting their trip. This option ranked first with a score of 3.65 out of 5. The scores for all answers are shown in Figure 7-2.

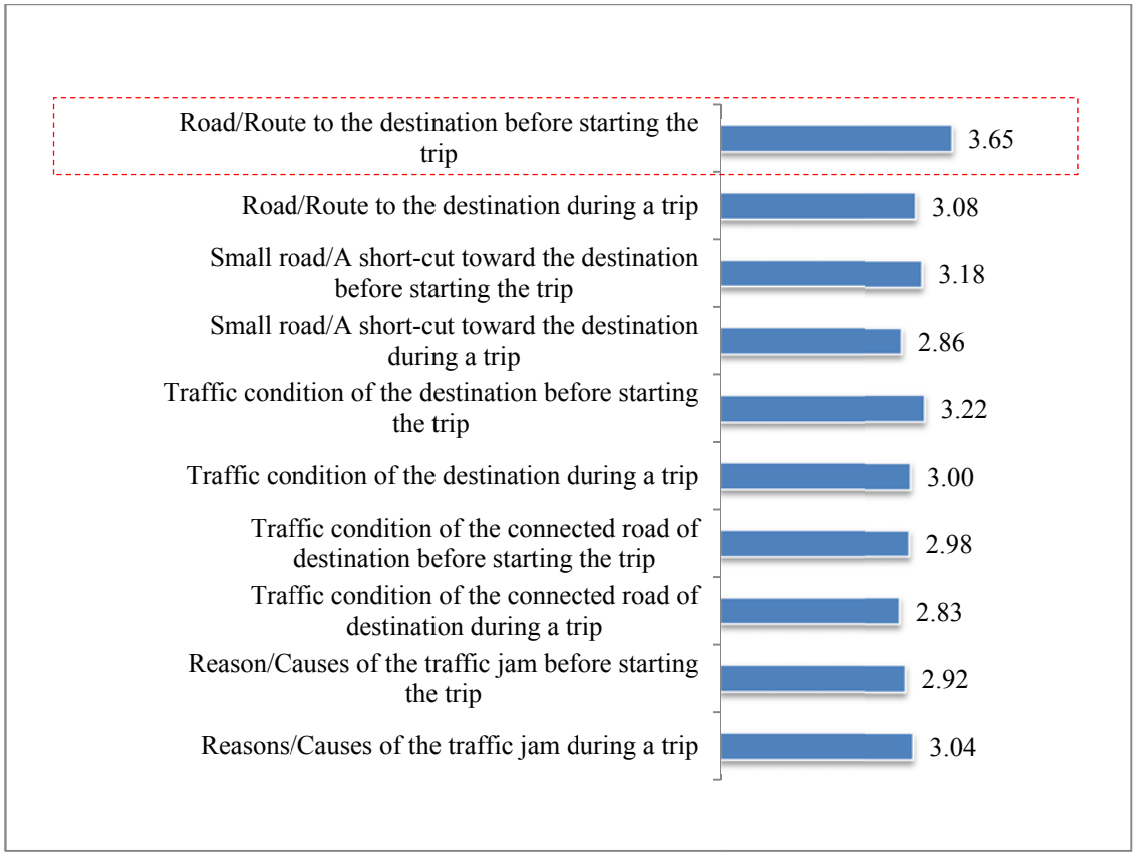



**Figure 7-2: Traffic information sought by users**

From the information presented in Figure 7-2, when grouped by types of information, the majority of Bangkok road users usually search for the road/route to the destination (ranked first with a score of 3.37). The traffic condition of the destination is ranked

second (3.11), followed by small road/short-cut to the destination (3.02), reasons/causes of a traffic jam (2.98) and traffic condition of a connected road of the destination (2.91) respectively.

Furthermore, when organizing results by time, the majority of respondents prefer to search for traffic information before starting the trip (3.19) follow by searching for information during the trip (2.96).

### 7.1.3 Route Planning in Advance and Alternative Route Selection

Figure 7-3 shows the course of action Bangkok road users would take if they knew in advance that a road in their route was congested. According to the results, 86.3% of the respondents would "decide to choose an alternative route instead of the congested one". The result is illustrated in Figure 7-3.

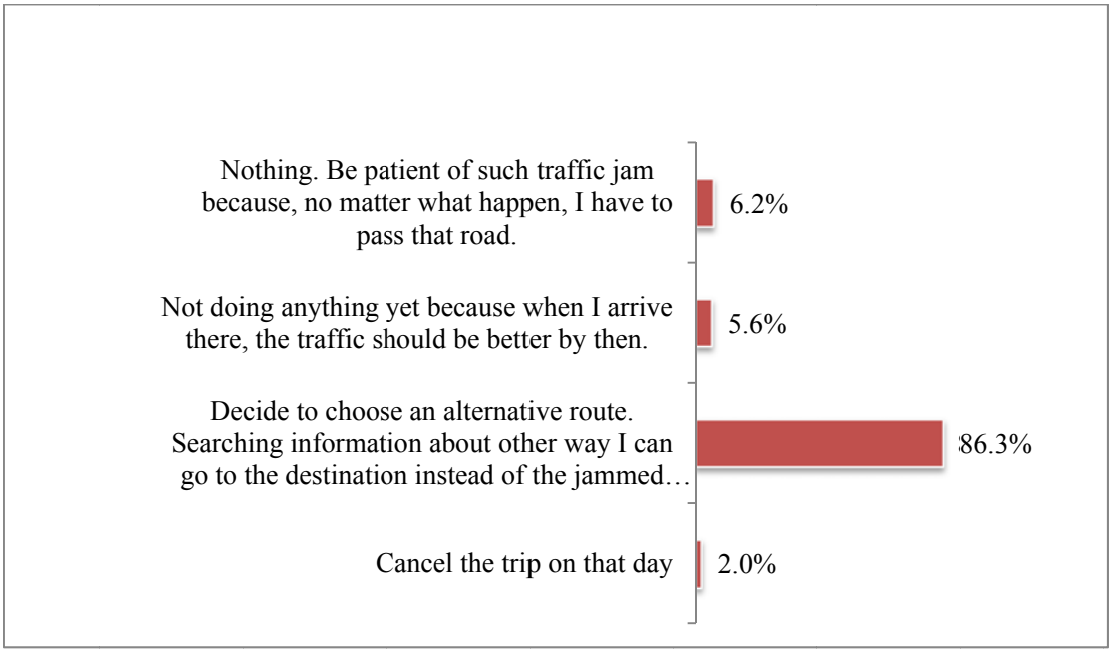



**Figure 7-3: Route planning in advance and alternative route selection**

### 7.1.4 Preferred Sources of Traffic Information

To study road users' preferred source of traffic information, we asked the question "Imagine you are able to access and know the traffic condition from the following

sources without any restriction. To what extent do you think the information from those sources will be useful for your driving?" The respondents give a score from 1 to 5, with 1 indicating 'Not at all useful' and 5 indicating 'Very useful'. The average score of each item was calculated and is shown in Figure 7-4.


**Figure 7-4: Preferred information sources when no restrictions apply**

When grouping responses by preferred information source, the majority of respondents ranked receiving traffic information via applications on mobile devices

first with a score of 3.87. Radio was ranked second (3.74). Information from websites (3.53) and electronic boards (3.34) were ranked third and the forth respectively.

From Figure 7-4, it is noticeable that when no limitations are placed on accessing information sources, respondents would prefer to access traffic information from applications on mobile devices. Although radio is popular, information from websites is also important. In the following discussion, we thus further analyze the traffic information that should be included if mobile applications are the preferred source of traffic information in the future. We then analyze the traffic information that should be included if websites are the preferred source of traffic information.

To further investigate the traffic information that Bangkok road users would like to access when using applications on mobile devices, we re-categorized our data by selecting only the responses that related to mobile devices. We also selected data based on responses to the question "How often do you search for each choice of traffic information provided?" where, using a Likert scale [40, 151], respondents could choose between never (1), rarely (2), sometimes (3), often (4) and every time (5). We choose only answers with responses from 3 to 5 because these are the respondents interested in traffic information via mobile devices while answers with responses of 1 or 2 indicted that these respondents were not interested in accessing traffic information via mobile applications. The Likert scale [40, 151] is commonly used in survey research to measure respondents' attitude to a particular question. The observed counts for a one-way frequency table of the selected data are shown in Table 7-1.

After re-categorization, we used the chi-square goodness of fit test to determine whether the association of traffic information search (road/route to the destination, small road/short-cut to the destination, traffic condition of the destination, traffic condition of connected road of destination and reasons/causes of the traffic jam) and the source of traffic information (in this case, applications on mobile devices) is in the expected ratio specified in our null hypothesis.

**Table 7-1: Observed counts for searching different types of traffic information via applications on mobile devices**

| How often do you search for the following information? | Sometimes | Often | Everytime | Total |
|---|---|---|---|---|
| Road/route to the destination | 255 | 332 | 176 | 763 |
| Small road/short-cut to the destination | 204 | 178 | 75 | 457 |
| Traffic condition at the destination | 317 | 288 | 154 | 759 |
| Traffic condition of roads connected to the destination | 243 | 127 | 67 | 437 |
| Reason/causes of a traffic jam | 211 | 124 | 71 | 406 |

From Table 7-1, if we believe that when Bangkok road users search for traffic information using a mobile application, they will search for the following information in a ratio of 2:1:2:1:1 at a significance level of 0.01:

- road/route to the destination
- small road/short-cut toward the destination
- traffic condition at the destination
- traffic condition of roads connected to the destination
- reason/causes of a traffic jam

then, for our chi-square goodness of fit test, the hypotheses take the following form:

$H_0$ : $P_1 : P_2 : P_3 : P_4 : P_5 = 0.284 : 0.144 : 0.284 : 0.144 : 0.144$

$H_1$ : The data are not consistent with a specified distribution.

The results from SPSS for the selected data presented in Table 7-1 are shown in Table 7-2 and Table 7-3.

**Table 7-2: Chi-square goodness of fit test (applications on mobile devices)**

| | Observed N | Expected N | Residual |
|---|---|---|---|
| Road/route to the destination | 763 | 801.5 | -2.3 |
| Small road/short-cut toward the destination | 457 | 406.4 | 97.7 |
| Traffic condition at the destination | 759 | 801.5 | -95.3 |
| Traffic condition of roads connected to the destination | 437 | 406.4 | -32.3 |
| Reason/causes of a traffic jam | 406 | 406.4 | 18.7 |
| Total | 2,822 | | |

**Table 7-3: Test statistics (applications on mobile devices)**

| | |
|---|---|
| Chi-Square | 12.711 |
| df | 4 |
| Asymp. Sig. | .017 |

As shown in Table 7-3, we can report these results as chi-square = 12.711, df = 4 and asymptotic significance = 0.017. The asymptotic significance (or *p*-value) is higher than the significance level (0.01) so thus we can accept $H_0$. We can conclude that if Bangkok road users use traffic reports on mobile applications, they will search for road/route to the destination : small road/short-cut toward the destination : traffic condition at the destination : traffic condition of roads connected to the destination : reason/causes of a traffic jam in the ratio 2:1:2:1:1 at a significance level of 0.01.

The results of the analysis suggest that mobile applications that provide traffic information to Bangkok road users should include the "Road/route to the destination" and "Traffic condition at the destination" because these are preferred at a ratio of 2 to 1 to the other options. Even if all the different types of information can be provided to road users without restriction, the placement of the "road/route to the destination" and "traffic condition at the destination" information options should be such that these options can be readily and easily accessed by road users.

We also analysed the data to further investigate the traffic information that Bangkok road users would like access when using websites that report traffic information. We selected data based on responses to the question "How often do you search for each

choice of traffic information provided?" where, using a Likert scale respondents could choose between never (1), rarely (2), sometimes (3), often (4) and every time (5). We choose only answers with responses from 3 to 5 because these are the respondents interested in traffic information via websites, while answers with responses of 1 or 2 indicated that these respondents were not interested in accessing traffic information via websites. The observed counts for a one-way frequency table of the selected data are shown in Table 7-4.

After re-categorization, we used the chi-square goodness of fit test to determine whether the association of traffic information search (road/route to the destination, small road/short-cut to the destination, traffic condition of the destination, traffic condition of connected road of destination, and reason/causes of a traffic jam) and the source of traffic information (in this case, websites) is in the expected ratio specified in our null hypothesis.

**Table 7-4: Observed counts for searching for different types of traffic information via websites**

| How often do you search for the following information? | Sometimes | Often | Everytime | Total |
|---|---|---|---|---|
| Road/route to the destination | 380 | 356 | 136 | 872 |
| Small road/short-cut to the destination | 198 | 208 | 87 | 493 |
| Traffic condition at the destination | 372 | 298 | 148 | 818 |
| Traffic condition of roads connected to the destination | 277 | 127 | 84 | 488 |
| Reason/causes of a traffic jam | 241 | 134 | 79 | 454 |

From Table 7-4, if we believe that when Bangkok road users search for traffic information on websites, they will search for the following information in a ratio of 2:1:2:1:1 at a significance level of 0.01:

- road/route to the destination
- small road/short-cut toward the destination
- traffic condition at the destination
- traffic condition of roads connected to the destination
- reason/causes of a traffic jam

then, for a chi-square goodness of fit test, the hypotheses take the following form:

$H_0$ : $P_1 : P_2 : P_3 : P_4 : P_5 = 0.284 : 0.144 : 0.284 : 0.144 : 0.144$

$H_1$ : The data are not consistent with a specified distribution.

The results from SPSS for the selected data presented in Table 7-4 are shown in Table 7-5 and Table 7-6.

**Table 7-5: Chi-square goodness of fit test (websites)**

| | Observed N | Expected N | Residual |
|---|---|---|---|
| Road/route to the destination | 872 | 887.5 | -1.7 |
| Small road/short-cut toward the destination | 493 | 450.0 | 24.7 |
| Traffic condition of the destination | 818 | 887.5 | -94.3 |
| Traffic condition of roads connected to the destination | 488 | 450.0 | 17.3 |
| Reason/causes of a traffic jam | 454 | 450.0 | -28.7 |
| Total | 3,125 | | |

**Table 7-6: Test statistics (websites)**

| | |
|---|---|
| Chi-Square | 13.067 |
| df | 4 |
| Asymp. Sig. | .015 |

As shown in Table 7-6, we can report these results as chi-square = 13.067, df = 4 and asymptotic significance = 0.015. As the asymptotic significance (or *p*-value) is higher than the significance level (0.01), we can accept $H_0$. We can thus conclude that if Bangkok road users access traffic information via traffic information websites, they will search for road/route to the destination: small road/short-cut toward the destination : traffic condition at the destination : traffic condition of roads connected to the destination : reason/causes of a traffic jam in the ratio 2:1:2:1:1 at a significance level of 0.01.

The results of the analysis result suggest that websites that provide traffic information to Bangkok road users should include the "road/route to the destination" and "traffic condition at the destination" because these are preferred at a ratio of 2 to 1 to the other options. Even if all the different types of information can be provided to road users without restriction, the placement of the "road/route to the destination" and "traffic condition at the destination" information options should be such that these options can be readily accessed by road users. It is noticeable that the results of the analysis for both mobile applications and websites are very similar.

### 7.1.5 Implications for Traffic Report Services

The effectiveness of traffic information distribution depends (partly) on a well-designed traffic information dissemination system. Generally, a traffic report service consists of three functions: data collection, data processing and information dissemination. A traffic report service is one of the keys to solving a traffic congestion problem. The analysis and results shown in this section can be used to help guide the development of traffic dissemination systems.

Our earlier analysis indicated that a traffic report service should assist users by helping them find a route to their destination (including small roads or shortcuts) and by informing them of traffic conditions at their destination. Road users also indicated they would like to know traffic conditions on connecting roads (see Section 7.1.2). Extrapolating from these results, we can assume that Bangkok road users are interested in the traffic conditions of minor roads in order plan their trip better. The CATE framework is able to support this requirement as the framework can be applied to minor roads lacking sensor infrastructure. Our CATE framework can thus help fulfil the Bangkok road users' information needs.

At the time of this research, more road users in Bangkok receive traffic information from radios than through mobile applications. However, since the 3G spectrum auction in Thailand in the last quarter of 2012, the number of smart phones and tablets, and the amount of mobile data usage, have increased substantially and they are continuing to grow rapidly [149, 150]. Preparing traffic information so that it can be disseminated through mobile devices is a logical step to meet this growing trend. The results from Section 7.1.4 show that most users would prefer to receive traffic

information via applications on mobile devices if they are able to choose without restriction.

The statistical analysis in Section 7.1.4 indicates that if mobile applications or websites are created to provide traffic information to Bangkok road users, the application or the website should include "road/route to the destination" and "traffic condition at the destination" as parts of the main features. However, if all types of traffic information (for example, road/route to destination, small road/short-cut toward destination, traffic condition of destination, traffic condition of roads connected to destination and reason/causes of a traffic jam) can be provided, then priority should be placed on users being able to readily access "road/route to destination" and "traffic condition of the destination" more than other features.

These results should be taken into consideration and used to help guide the development of traffic report services.

## *7.2 The Potential Use of Social Networks for TIS in Bangkok*

The advent and development of online social networks has been one of the most exciting events in this decade. Many online social networks such as Twitter, LinkedIn, Facebook and Google+ have become increasingly popular. Such social networks are rich in content and they typically contain content that can be leveraged for analysis.

This investigation into the role of social networks in traffic systems was motivated by our belief that the traffic data extracted from social networks of Bangkok road users might be able to be used as a context attribute for calculating the traffic congestion degree. Although investigating this proposition and its implementation in depth is left for future work, we briefly consider its feasibility here.

Many approaches to utilize social networks in traffic reports have been proposed in recent years. For example, the National Electronics and Computer Technology Centre of Thailand created its own Twitter account named @traffy to convey traffic updates to road users and to receive information from the public [152]. Twitter has also been used by TMC Polda Metro Jaya Indonesia to spread traffic news. This was achieved

by using natural language processing to extract traffic information from tweets and to reflect that data onto a map [38]. In addition, an attempt to harvest road traffic information in Indonesia from virtually indeterminate sources of data on Twitter timelines for real time mapping was achieved in [39].

Even though messages can be distributed through social networks to assist with traffic information dissemination, and researchers have proposed methods to harvest traffic information from social networks, those works were not based on, nor developed for, Bangkok road users. Improving the Bangkok traffic report service in the face of current technology and social network patterns has yet to be explored. To rectify this omission and improve the existing system, we began by designing a survey to discover how Bangkok road users use social networks and the potential of using social networks to provide useful traffic data in Bangkok. The results of our analysis (reported below) indicate that it is feasible to give social networks a role to play in Bangkok's TIS.

### 7.2.1 Frequency of Social Network Access

When asking how often respondents access social networks, the results show that the majority of drivers (77.1%) spend time on a social network everyday (illustrated in Figure 7-5).

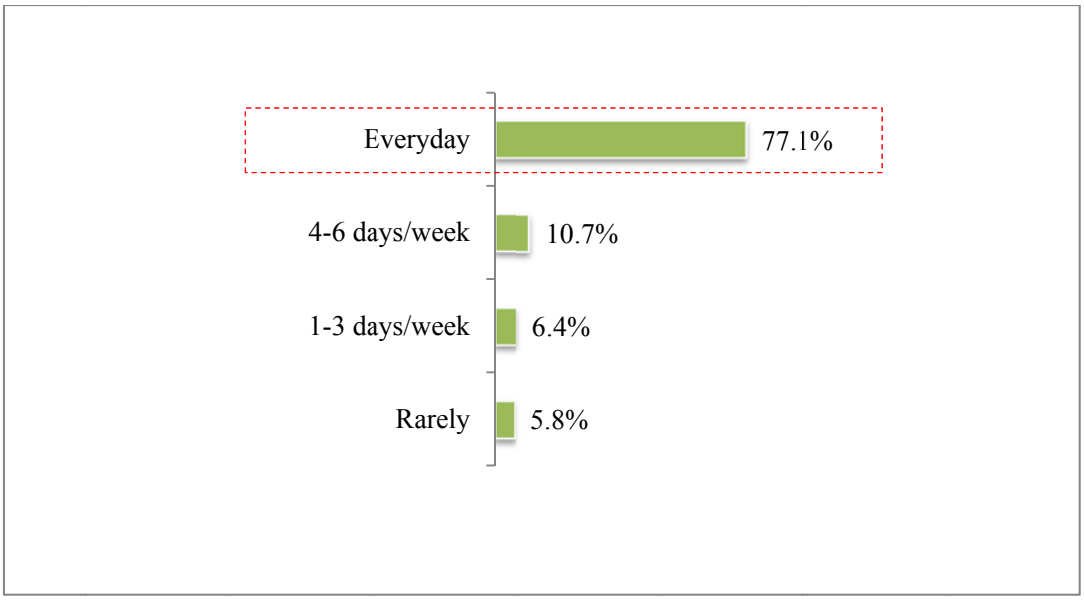



**Figure 7-5: Frequency of social network access**

### 7.2.2 Frequency of Mentioning Traffic Conditions in Social Networks

For those who use social networks, we asked how often they create posts that mention traffic conditions. The result is illustrated in Figure 7-6.

Although about half of the drivers (54%) rarely mention traffic conditions, nearly 40% of them mention traffic conditions often (a combination of 'quite often' and 'very often' in the results).

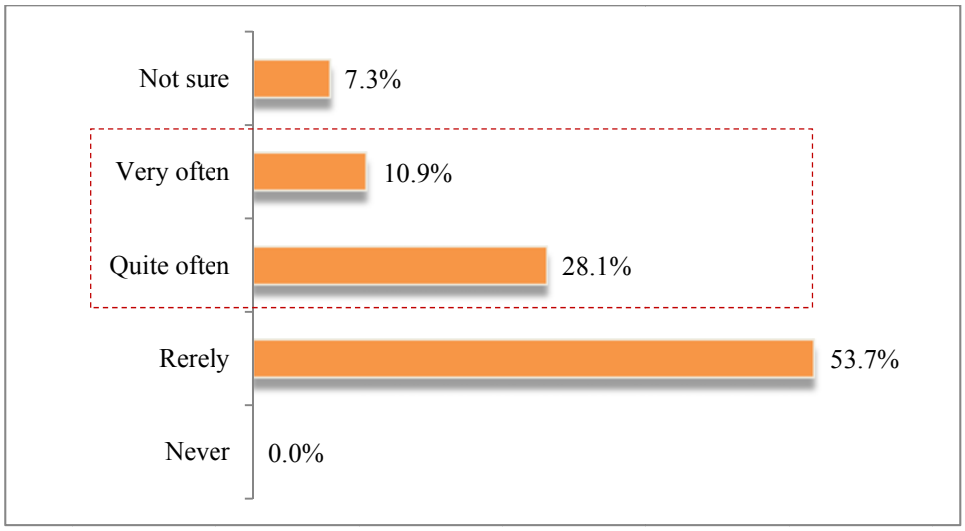



**Figure 7-6: Frequency of mentioning traffic conditions in social networks**

### 7.2.3 Traffic Content Mentioned in Social Networks

For those who mention traffic conditions on social networks, we asked the question "if you mention traffic conditions in your social network post, how do you mention it?" to discover what information related to traffic they include. The result is summarized in Figure 7-7.

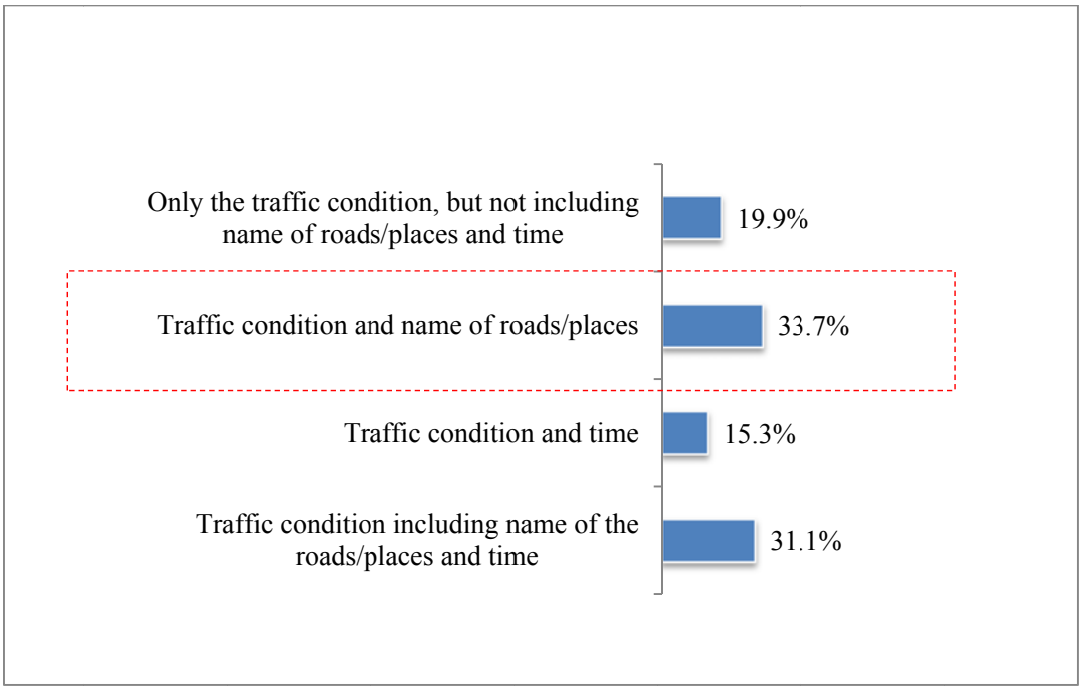



**Figure 7-7: Traffic content mentioned in social networks**

It can be seen that most traffic content mentioned by respondents in social networks refers to traffic conditions at a particular place (33.7%). The traffic condition, including the place and time the traffic condition occurs, is ranked second (31.1%). Mentioning only the traffic condition (19.9%) or the traffic condition including time but not place (15.3%) is ranked third and fourth respectively. These results are discussed in Section 7.2.5.

### 7.2.4 Relationship between results

To facilitate interpretation of the results and to show the relationship between variables, we cross-tabulated our results. Table 7-7 describes the relationship between the frequency of social network use of Bangkok road users and the frequency that road users mention traffic conditions when they create posts on social networks. Table 7-8 describes the relationship between the frequency with which road users mention traffic conditions when creating posts on social networks and how context is included within the posts if traffic conditions are mentioned.

**Table 7-7: Cross tabulation between the frequency of accessing social networks and the frequency of mentioning traffic conditions**

| | | | Mention_traffic_cond | | | | |
|---|---|---|---|---|---|---|---|
| | | | Rarely | Not sure | Many time | Often | Total |
| Use Social Network | Rarely | Count | 30 | 8 | 8 | 4 | 50 |
| | | % within Use Social Network | 60.0% | 16.0% | 16.0% | 8.0% | 100.0% |
| | | % within Mention_traffic_cond | 6.5% | 12.7% | 3.3% | 4.3% | 5.8% |
| | | % of Total | 3.5% | .9% | .9% | .5% | 5.8% |
| | 1-3 days/week | Count | 39 | 6 | 7 | 3 | 55 |
| | | % within Use Social Network | 70.9% | 10.9% | 12.7% | 5.5% | 100.0% |
| | | % within Mention_traffic_cond | 8.4% | 9.5% | 2.9% | 3.2% | 6.4% |
| | | % of Total | 4.5% | .7% | .8% | .3% | 6.4% |
| | 4-6 days/week | Count | 69 | 8 | 13 | 2 | 92 |
| | | % within Use Social Network | 75.0% | 8.7% | 14.1% | 2.2% | 100.0% |
| | | % within Mention_traffic_cond | 14.9% | 12.7% | 5.4% | 2.1% | 10.7% |
| | | % of Total | 8.0% | .9% | 1.5% | .2% | 10.7% |
| | Everyday | Count | 324 | 41 | 214 | 85 | 664 |
| | | % within Use Social Network | 48.8% | 6.2% | 32.2% | 12.8% | 100.0% |
| | | % within Mention_traffic_cond | 70.1% | 65.1% | 88.4% | 90.4% | 77.1% |
| | | % of Total | 37.6% | 4.8% | 24.9% | 9.9% | 77.1% |
| Total | | Count | 462 | 63 | 242 | 94 | 861 |
| | | % within Use Social Network | 53.7% | 7.3% | 28.1% | 10.9% | 100.0% |
| | | % within Mention_traffic_cond | 100.0 | 100.0% | 100.0% | 100.0% | 100.0% |
| | | % of Total | 53.7% | 7.3% | 28.1% | 10.9% | 100.0% |

**Table 7-8: Cross tabulation between the frequency of mentioning traffic conditions and the type of context included when mentioning traffic conditions**

| | | | How do you mention? | | | | |
|---|---|---|---|---|---|---|---|
| | | | Only the traffic cond. | Traffic condition and name of roads | Traffic condition and time | Traffic cond. and name of roads+time | Total |
| Mention about traffic condition | Rarely | Count | 109 | 162 | 70 | 121 | 462 |
| | | % within Mention_traffic_cond | 23.6% | 35.1% | 15.2% | 26.2% | 100.0% |
| | | % within How do you mention? | 63.7% | 55.9% | 53.0% | 45.1% | 53.7% |
| | | % of Total | 12.7% | 18.8% | 8.1% | 14.1% | 53.7% |
| | Not sure | Count | 13 | 23 | 11 | 16 | 63 |
| | | % within Mention_traffic_cond | 20.6% | 36.5% | 17.5% | 25.4% | 100.0% |
| | | % within How do you mention? | 7.6% | 7.9% | 8.3% | 6.0% | 7.3% |
| | | % of Total | 1.5% | 2.7% | 1.3% | 1.9% | 7.3% |
| | Many time | Count | 37 | 78 | 38 | 89 | 242 |
| | | % within Mention_traffic_cond | 15.3% | 32.2% | 15.7% | 36.8% | 100.0% |
| | | % within How do you mention? | 21.6% | 26.9% | 28.8% | 33.2% | 28.1% |
| | | % of Total | 4.3% | 9.1% | 4.4% | 10.3% | 28.1% |
| | Often | Count | 12 | 27 | 13 | 42 | 94 |
| | | % within Mention_traffic_cond | 12.8% | 28.7% | 13.8% | 44.7% | 100.0% |
| | | % within How do you mention? | 7.0% | 9.3% | 9.8% | 15.7% | 10.9% |
| | | % of Total | 1.4% | 3.1% | 1.5% | 4.9% | 10.9% |
| Total | | Count | 171 | 290 | 132 | 268 | 861 |
| | | % within Mention_traffic_cond | 19.9% | 33.7% | 15.3% | 31.1% | 100.0% |
| | | % within How do you mention? | 100.0 | 100.0% | 100.0% | 100.0% | 100.0% |
| | | % of Total | 19.9% | 33.7% | 15.3% | 31.1% | 100.0% |

### 7.2.5 Implications of Results for the Potential Use of Social Networks in Bangkok's TIS

Analysis of results indicate that the substantial number of social network users in Bangkok make it feasible to take advantage of social networks for Bangkok's TIS. The majority of Bangkok drivers (77.1%) regularly spend time on social networks in everyday life and approximately 40% of these drivers mention traffic conditions. Although about 54% of Bangkok road users rarely mention traffic conditions, nearly 40% of them mention traffic conditions often in their posts on social networks, while 10.9% mention traffic conditions very often and 28.1% mention traffic conditions quite often. About 33% of those mentioning traffic conditions include locations in their posts.

From Table 7-7, it is noticeable that 28.1% of road users who normally use social networks mention traffic condition many times. Furthermore, 48.4% use social networks every day but rarely mention traffic conditions over social networks.

Table 7-8 indicates that many of the respondents who regularly mention traffic conditions in social networks will include the *traffic condition, location,* and *time* (36.8%) in their posts. This result indicates that social networks are potentially a valuable resource for gleaning data relevant to TIS. Consequently, the traffic data harvested from social networks could be taken as an influential context attribute for our adaptive real time traffic congestion estimation system to enhance the competency and accuracy of the framework.

One possible approach to extract traffic data from social networks is to create a specific social network account, such as a Twitter account, to report and receive the traffic information from distributed users. This method is easier to implement than collecting all content distributed over all social networks because most of the content in the especially created social network account would only be traffic information, thus narrowing the search effort and minimizing associated costs. For this approach to work, a critical mass of road users would be required to ensure useful traffic data can both be supplied to and enough obtained from the system. Therefore, such account should be promoted to gain a great number of engaged users when deploying this approach. It is noteworthy that we received sizable feedback in a short period of time

after distributing our online questionnaire through social network accounts with over 400,000 followers. Therefore it is significant to let numerous users participated if we decide to use social network as part of traffic report services.

Some limitations are associated with gaining traffic context attributes from social networks. First, traffic information gained from a social network may be obsolete and inaccurate because it might not be posted at the exact moment and place of the traffic condition. Second, even if posted immediately, the traffic information may only be valid for a short amount of time and may become obsolete soon after it is posted. Third, the creditability of information exchanged through social networks maybe low. Fourth, traffic data posted on social networks is only posted intermittently, at the whim of the social network user. Finally, many of the messages about traffic information posted on social networks are ungrammatical [39]. Natural language processing, syntactic analysis and social media data mining techniques shall be applied to garner traffic data from social networks [153].

Because of these limitations, from our point of view, traffic data extracted from a social network site is more suitable for use in a supporting, rather than core, role. For example, traffic data from a social network site could be reserved for use as a context attribute in the CATE framework only when other influential context attributes cannot be acquired at run time. Alternatively, traffic data extracted from a social network could be used to calculate the additional confidence factor to assist in confirming the calculated inferred traffic congestion degree, rather than relying on it as the main input for traffic estimation.

## *7.3 Chapter Summary*

In this chapter, we presented our statistical analysis of the collected data, derived from a web-based survey, on the traffic information usage of Bangkok road users and the potential role of social networks for TIS in Bangkok, Thailand. The results reported in this chapter contribute to our knowledge of Intelligent Transportation Systems (ITS) and to potential traffic management improvement, not only for Bangkok, but also for other cities that share characteristics similar to Bangkok.

Analysis of our results showed that users value information regarding journey route and road conditions relevant to that route. The results also indicate that, in addition to radio, road users would prefer to access this information via applications on mobile devices and through websites. Our investigation also showed that road drivers contribute traffic information through social networks, and that these, while not without their limitations, are used by a sufficient number of drivers that traffic data extracted from social networks could be used to make a meaningful contribution to Bangkok's TIS. In terms of the CATE framework, traffic information extracted from a social network site could be used as a context attribute when other contexts are unavailable, and could also be used as the additional confident factor to confirm the calculated inferred traffic congestion degree. The information arising from our investigation in this chapter may offer some insights and thus be beneficial to and guide the development of future TIS and the improvement of existing systems.

The results discussed in this chapter might also be useful to other researchers who are interested in researching traffic information and traffic report services. Although our study was based on Bangkok data, it may be applicable to other cities that share similar road infrastructure and information issues. In the next and final chapter, we consider in more detail the contributions made by this research.

# Chapter 8 Conclusion and Future Works

This chapter provides a close to this thesis and addresses the final stage of the design science research method: conclusion. At this stage, the results and knowledge gained are consolidated and made explicit and are thereby made available for further research. This final chapter begins with a research summary by presenting key findings and outcomes relevant to Traffic Information Systems (TIS) in the Intelligent Transportation System (ITS) domain. The chapter highlights our research contribution and summarizes an adaptation of Hevner et al.'s [33] design science research guidelines to evaluate the research quality. We suggest directions for potential future research and then finish with the chapter summary.

## *8.1 Research Summary*

This research was undertaken to find a solution that addresses the problem of unreliable sensory traffic data and subsequent intermittent availability of traffic information to drivers in Bangkok. To overcome this issue, we sought a solution that utilized available contexts in the environment. We conducted the research by following the design science research methodology and adopting the design science research cycle proposed by Hevner et al. [32, 33] as its overarching framework. We followed the design science research steps proposed by Vaishnavi and Kuechler [34, 35] for the research process. The outcome of following this research process, and our solution to the above problem, was the Context-Aware Traffic Congestion Estimation Framework to Overcome Missing Sensory Data (the CATE framework).

The design of our CATE framework drew upon several different knowledge domains, including ITS, TIS, pervasive computing, data mining and machine learning. In particular, because traffic congestion has been a problematic experience for commuters in metropolitan areas, in Chapter 2, we reviewed the research into TIS. A TIS, which is a component of an ITS, can play a significant role in improving traffic

congestion problems. Sensor data is fed into a TIS and processed, with the output - information about real time traffic - assisting drivers in planning their routes to avoid traffic congestion. Most existing traffic state estimation approaches rely on data from sensors of an observed road segment. However, sensors may become unavailable due to damage, communication breakdown or poor environmental conditions, or some roads may lack installed sensors. Consequently, sensor data may be lacking, rendering traffic information intermittent or unavailable. The estimation of traffic conditions when sensory traffic data is lacking does not appear in most works and other types of data sources in the environment are rarely taken into account when estimating traffic conditions.

To overcome this paucity, and to solve the issue of poor quality traffic information, in Chapter 4 of this thesis we proposed a Context-Aware Traffic Congestion Estimation Framework to Overcome Missing Sensory Data. Unlike most existing methods, our approach, when sensor data is unavailable, is able to utilize available external contextual data instead. The CATE framework uses a machine learning algorithm to build inference models and then employs the available context attributes in run time to infer the traffic congestion degree when the sensory traffic data of observed road segments is not available. To deal with the possibility of changes to traffic situations that may make predictions less accurate, the CATE framework incorporates a built-in relearning function that can be used to improve the accuracy of inference models over time.

We demonstrated the feasibility and efficiency of our proposed framework through experimentation in Chapter 5 (Evaluation I). We also compared the accuracy of the framework with other approaches and showed that the CATE framework performs better than the other approaches both in terms of accuracy and time. However, we also found that we can improve our proposed framework to reduce both processing time and costs (computing resources, time, human resources and effort) to enhance the overall quality of the framework.

The knowledge gained from the first evaluation, combined with further investigation into road users' perceptions of the factors influencing traffic conditions in Bangkok, were used to construct a new list of influential context attributes: the reduced set of influential context attributes (*RS*). This new set of attributes was then used to improve

the design artefact. In Chapter 6, we presented the end result of this iterative process of artefact development and evaluation: the refined CATE framework. In Chapter 6 we also conducted further evaluation to assess the feasibility, efficiency and improvement of our final artefact compared to the initial framework, and to demonstrate its superior performance, in terms of accuracy, over traditional approaches. Figure 8-1 summarises this iterative process. The organization of this thesis, including the overall process of our research and its outcome, can be found in Figure 3-4.

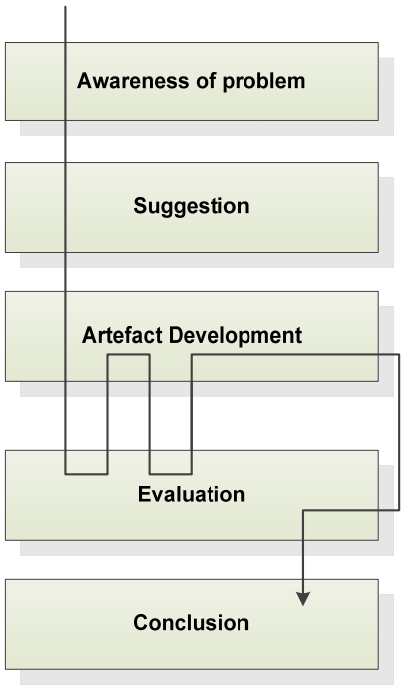



**Figure 8-1: Research process summary**

In addition to designing the CATE framework, we also performed further research to derive recommendations to facilitate the implementation of TIS and traffic report services in Bangkok. This research involved statistical analysis of quantitative results from a web-based survey. The survey sought responses about the traffic information needs of Bangkok drivers and their preferred media channels to receive this information. The survey also aimed to study the potential use of social networks in Bangkok's TIS with the view of determining whether this data could be used to

improve both Bangkok's TIS and traffic report services. This study was reported in Chapter 7.

## 8.2 Research Findings and Contribution

This research is based on the design science paradigm. It aims to develop an artefact that will enable a TIS to tolerate uncertain situations due to unreliable sensory traffic data and be resilient to scarce input. We attempt to solve the existing problem of insufficient traffic information resulting from missing sensory data, whether due to poor weather conditions or resource constraints. Guidelines to improve TIS implementation are also provided. Our CATE framework has the positive characteristics of *adaptability* as the framework can adapt itself to different situations based upon the availability of contextual information in the environment and *extensibility* as new road segments can be easily added to the CATE system.

The relearning function included in our CATE framework allows our framework to cope with the possibility of changes to the traffic situation that may make traffic predictions less accurate. The CATE framework can improve in accuracy automatically as actual data is collected. In addition, the relearning function allows the CATE framework to cope with sparse historical data. Our framework could thus also be applied to road segments with newly installed traffic sensors.

The sets of factors that influence traffic congestion are also identified in our research, which is useful for TIS implementation. In addition, *RS,* which we used to improve our artefact, and the process used to select *RS*, can be used to identify the context attributes that warrant investment or, alternatively, should be ignored. This can be useful for traffic management sector. The research also makes contributions to improve the implementation of Bangkok's TIS through the survey and subsequent statistical analysis based on the perceptions of Bangkok's road users.

This thesis makes theoretical, methodological and practical contributions, through the conceptualization, development and evaluation of our CATE framework, to the ITS domain and community. The contributions of our research are explained, as follows.

### 8.2.1 Contribution to the ITS

This research is an exercise in using the design science approach to produce a designed artefact that can solve an existing problem. This research provides a better understanding of the use of design science as a research methodology for solving problems in the ITS domain. This research shows that design science can address and solve ongoing theoretical and practical issues in ITS. Thus, it contributes to the theory of doing research in ITS.

Furthermore, this research contributes novel findings for the TIS and ITS community. Our research has added to the knowledge base of the ITS community by introducing a new, adaptive, context-aware framework that can overcome the problem of missing sensory traffic data by also collating acquirable surrounding context data instead of solely relying on sensory data. This enables traffic report services to provide complete and continuous traffic information even though the acquirable influential contexts attributes is sparse. Our approach is thus different from most existing approaches. To the best of our knowledge, no TIS in Thailand provides a solution to the problem of missing sensory traffic data, whether due to poor environmental conditions or resource constraints. If our proposed framework is adopted, it will enable a TIS to provide more complete information and thus benefit road users in route planning.

In addition, *RS* not only improves the CATE framework, but it also adds new knowledge to the ITS domain by defining a minimum sufficient set of factors required to estimate traffic conditions. This is also useful to anyone who desires to implement a TIS in Bangkok (and in other cities that have characteristics similar to Bangkok). Thus, it contributes both theoretically and practically.

Moreover, the method we conduct to obtain *RS* can be applied as a tool by traffic management organizations to analyse and identify the context attributes that justify investment or, alternatively, should be ignored. This shows both methodological and practical contribution.

Through the survey based on the perceptions of Bangkok's road users and subsequent statistical analysis, the thesis also makes contributions to the development of TIS' reporting systems, as can be seen in Chapter 7. Our online survey and statistical

analysis of Bangkok's road users' traffic information needs can guide future implementations of traffic report services for TIS. The guidelines for designing mobile applications and websites to provide traffic information (developed from our survey and statistical analysis) are not only useful to facilitate traffic report services as an extension of our proposed framework, but they can also be used to improve other existing TIS. Furthermore, the findings from our study on the utilization of social network content also contribute to TIS implementation. Our findings are also applicable to existing TIS to improve efficiency.

As stated in [33], the criteria for assessing contribution focuses on representational fidelity and implementability. The CATE framework is applicable to existing TIS. Our evaluations have proved that our artefact is feasible and implementable with minimal complexity.

In summary, this thesis makes theoretical, methodological and practical contributions to the ITS domain in many aspects.

### 8.2.2 Contribution to Community

The outcome of this thesis can help improve traffic dissemination and traffic report services, thus improving people's quality of life. Users can make better decisions when route planning to avoid jammed roads, yielding a benefit in time and fuel savings. Our research also contributes to community by producing a framework that has the potential to make a positive difference to road users by improving traffic information creation and dissemination. Moreover, our survey result (presented in Section 6.2 in Chapter 6) shows that over 94% of drivers in Bangkok value knowing traffic information. This result confirms the usefulness of our artefact.

## *8.3 Quality Assessment of the Research*

In order to assess the quality of the current research, in this section we use the seven guidelines for judging design science research quality suggested by Hevner et al. [33]. The purpose of Hevner et al.'s seven guidelines is to assist researchers to understand the requirements for effective design science research. This section presents an

adaptation of Hevner et al.'s seven guidelines as shown in Table 8-1 for conducting, evaluating and presenting design science research.

**Table 8-1: Guidelines for Judging Design-Science Research Quality by Hevner et al. [33]**

| Guideline | Description |
|---|---|
| Guideline 1: Design as an Artifact | Design science research must produce a viable artifact in the form of a construct, a model, a method, or an instantiation. |
| Guideline 2: Problem relevance | The objective of design science research is to develop technology-based approaches to important and relevant business problems. |
| Guideline 3: Design evaluation | The utility, quality, and efficacy of a design artifact must be rigorously demonstrated via well-executed evaluation methods. |
| Guideline 4: Research contributions | Effective design science research must provide clear and verifiable contributions in the areas of the design artifact, design foundations, and/or design methodologies. |
| Guideline 5: Research rigor | Design science research relies upon the application of rigorous methods in both the construction and evaluation of the design artifact. |
| Guideline 6: Design as a search process | The search for an effective artifact requires utilizing available means to reach desired ends while satisfying laws in the problem environment. |
| Guideline 7: Communication of research | Design science research must be presented effectively to both technology-oriented and management-oriented audiences |

### 8.3.1 Guideline One: Design as an Artefact

The result of design science research is a purposeful IT artefact created to address an important organizational problem. It must be described effectively, enabling its implementation and application in an appropriate domain [33]. The artefacts constructed in design science are innovations that define the ideas, practices, technical capabilities and products through which the analysis, design, implementation and use of information systems can be effectively and efficiently accomplished [154, 155]. The first guideline is concerned with the design science research producing a viable artefact in the form of a construct, a model, a method or an instantiation.

This research aimed to introduce a primary design artefact in the form of an adaptive framework that enables traffic information systems to tolerate uncertain, unreliable and at times, unavailable input data. The framework achieves this goal by utilizing current available external contexts, when data from sensors is lacking, to estimate the

level of traffic congestion. The research also highlights the sets of factors that influence traffic congestion. In addition, the research identifies a reduced set of influential context attributes. The process to arrive at this sufficient set of attributes is reproducible and can be used to decide the context attributes that merit investment. The efficiency, effectiveness and feasibility of the proposed method were confirmed through experiments in Chapter 5 and Chapter 6. In addition, the outcomes from the statistical analysis of the online survey concerning Bangkok's road users' perceptions offered suggestions for TIS implementation.

### 8.3.2 Guideline Two: Problem Relevance

The objective of design science research is to develop technology-based approaches to important and relevant business problems [33]. In our research, the relevance and importance of the problem were well identified. A literature analysis was conducted to examine the existing research in TIS to identify the key issues to be addressed as discussed in Chapter 2. The literature review on several works, which included the observation on existing systems in Bangkok, highlighted several issues.

1) Most existing traffic state estimation approaches rely on data from sensors that belong to an observed road segment. Other types of data sources in the environment are rarely taken into account.

2) A method to estimate the traffic condition in sensorless road segments or when sensory traffic data is lacking has yet to be proposed in most existing research.

3) To the best of our knowledge, at the time of our research, the TIS in Thailand does not provide a solution for missing sensory traffic data. Nor does it provide a solution for road segments lacking sensors.

4) Suitable influential context attributes for traffic condition estimation have rarely been studied. No investigations appear to have been conducted especially for Bangkok's traffic situation.

5) Suitable media channels and items for Bangkok traffic reports have rarely been studied.

Our design artefact is a technology-based solution to important and relevant TIS problems in the ITS domain.

### 8.3.3 Guideline Three: Design Evaluation

Evaluation is a crucial component of the research process. The evaluation method must be well selected and executed in order to demonstrate the utility, quality and efficacy of the design artefact [33, 156, 157].

Vaishnavi and Kuechler stated in [34] that to evaluate and validate an approach, and for the claims about the approach to be acceptable to the research community, certain patterns provide vehicles for the evaluation and validation. Those patterns are demonstration, experimentation, simulation, using metrics, benchmarking, logical reasoning and mathematical proofs.

According to the taxonomy provided by Hevner et al. [33], an artefact can be evaluated through the following methods:

- observational methods such as case study and field study;
- analytical methods such as static analysis, architecture analysis, optimization and dynamic analysis;
- experimental methods such as controlled experiment and simulation;
- testing methods such as functional (black box) testing and structural (white box) testing; and
- descriptive methods such as informed argument and scenarios.

The design artefacts of this research were rigorously evaluated through experimental methods. The experiment results from Evaluation I, along with further study into the perception of Bangkok's road users collected through web-based survey methods, were the foundations for the artefact's refinement. In Evaluation II, the refined framework was evaluated on its improvement, efficiency, and accuracy through the experimental method.

### 8.3.4 Guideline Four: Research Contributions

Most often the contribution of design science research is the artefact itself. The artefact must enable a solution to solve a problem. It may extend the knowledge base or apply existing knowledge in new and innovative ways. Criteria for assessing contribution focus on representational fidelity and implementability [33]. Our research makes theoretical, methodological and practical contributions to the ITS domain and makes contributions also to the general community.

The experiments in our evaluation phases demonstrated that the CATE framework is efficient and implementable with little complexity. The CATE framework was designed to be applicable to existing TIS. In addition, this thesis demonstrates the suitability of the design science research methodology for research in the ITS domain. Currently, only a small amount of ITS research employs a design science methodology. This thesis adds to this knowledge bank. Thus, the research methodology that we perform in this thesis has contribution to the research methodology in ITS domain.

Through the conceptualization and evaluation of our CATE framework, this thesis makes theoretical and practical contributions to the ITS domain and to community. Our research has added new knowledge to the ITS domain by introducing a new context-aware framework that can overcome missing sensory traffic data by additionally utilizing available surrounding context data. Furthermore, *RS* is not only useful for improving our CATE framework, but it also adds new knowledge to the ITS domain in terms of identifying the minimum sufficient set of factors required to estimate traffic conditions. *RS* is beneficial to TIS implementation in Bangkok (and perhaps to other cities similar to Bangkok). Moreover, the method we used to obtain *RS* can be applied as a tool to identify the influential context attributes that warrant investment or, alternatively, should be ignored.

The survey based on the perceptions of Bangkok's road users and subsequent statistical analysis make contributions to the development of TIS' reporting systems in the form of guidelines for designing mobile applications and websites. These are useful, not only for supporting TIS implementation as an extension of our proposed framework in the future, but they can also be used to improve existing TIS in other

locations. Furthermore, the findings from our investigation into the utilization of social media for TIS also contribute to improving the implementation of existing TIS.

Finally, our survey results indicate that 95% of drivers in Bangkok value traffic information. This confirms the usefulness of our artefact. Our research contributes to community by producing a framework that has the potential to make a positive difference to road users seeking traffic information.

### 8.3.5 Guideline Five: Research Rigor

Hevner et al. [33] suggest that design science research should rely upon the application of rigorous methods in both the construction and evaluation of the designed artefact. This research has adopted the design science research cycle proposed by Hevner et al. [32, 33] as its overarching framework and the design science research steps proposed by Vaishnavi and Kuechler [34, 35] for the research process. The designed artefact was designed and constructed using the existing knowledge base and theories of ITS, data mining, machine learning, pervasive computing and context awareness. Moreover, the evaluation of the design artefact was conducted rigorously using experimental methods.

### 8.3.6 Guideline Six: Design as a Search Process

Design science is inherently iterative. Design is essentially a search process to discover an effective solution to a problem. Progress is made iteratively as the scope of the design problem is expanded. The sixth guideline of Hevner et al. involves designing an effective artefact through an iterative search process. The search for an effective artefact requires utilising available means to reach desired ends while satisfying laws in the problem environment [33].

Figure 8-1 shows a summary of our iterative research process. This can be viewed as a search process, through artefact development and improvement, for the optimal solution. We utilize available theories and methods to design novel artefact that can solve a problem in TIS and consequently yield new knowledge that contributes to the ITS domain. The results and analysis from the online survey, combined with the knowledge gained from Evaluation I, were used to improve the design and create the final artefact.

### 8.3.7 Guideline Seven: Communication of Research

The seventh guideline requires the effective presentation of the design science research to research-, technology- and user-oriented audiences. Communication of research can happen both during and after the research process. During this research project, we have presented our interim findings at conferences and published sequence of publications as shown in our List of Publications. Feedback received when we presented our research at conferences was also used to improve our proposed artefact.

## *8.4 Future Research Direction*

Although our research covered a number of issues in depth, scope still exists for further exploration and improvement. In addition, as the research progressed, issues beyond the scope of the original project arose that are worthy of further investigation.

The CATE framework has a relearning function so that it can improve itself automatically over time as described in Section 4.4.3 in Chapter 4. The relearning task to rebuild inference models can be repeated at user defined time intervals or when new historical data is received. However, the optimal time period for relearning has yet to be identified, and could be the focus of future research.

The analysis of the survey resulted in guidelines for designing mobile applications and websites to help disseminate traffic information. A possible direction for future research is to extend the CATE framework taking these guidelines into consideration. Investigating how these outcomes could be used to improve other existing TIS could also be the subject of future research.

In addition, our study into the potential of using social networks for TIS in Bangkok indicates that it is feasible to both extract traffic data and to provide traffic information through social networks. The data extracted from social networks could be included as context attributes for the CATE framework to improve competency and accuracy in the future. Likewise, the traffic data posted in social networks could be used to calculate the confidence factor to confirm the level of correctness of the inferred traffic congestion degree of our proposed framework. Exploring these and similar possibilities could be the focus for future research.

## 8.5 Chapter Summary

This chapter reflects on the overall research process. The chapter highlights the research findings and contributions this thesis brought to the ITS domain and community. The adaptation of Hevner et al.'s seven guidelines [33] to evaluate design science research was discussed to demonstrate the quality of our research. Finally, the directions for future research, both for extending our proposed framework and for other existing TIS, were identified.

This research has produced a framework that has the potential to make a positive difference to researchers in the ITS domain and to Bangkok's road users. The results justify continuing research in this area in order to increase the body of scientific knowledge of the ITS domain, to provide practical support to those involved in managing and maintaining TIS, and to improve the quality of life of the wider community.

# References


[1] N. Tiwakorn, W. Rengchan and P. Apiwat, "Satisafaction of Drivers toward Intelligent Traffic Board in Bangkok " Master of Business Administration, Business Administration, Naresuan University, Pitsanulok, 2007.

[2] (2012). *10 monster traffic jams from around the world*. Available: http://www.bbc.co.uk/news/magazine-19716687.

[3] T. Nachaiwieng, "Economic Valuation of Traffic Congestion Costs in Bangkok," Master of Engineering, Civil Engineering, Chulalongkorn University, Bangkok, 2001.

[4] P. Kiukasam, "Resolutions of Traffic Problems in Bangkok Metropolitan: A Comparative Study Between Opinions of Traffic Policeman Posted at Police Station and those Posted at Police's Traffic Division," Master of Arts, Political Science, Ramkamheng University, Bangkok, 1995.

[5] A. A. Shah and D. Lee Jong, "Intelligent Transportation Systems in Transitional and Developing Countries," *Aerospace and Electronic Systems Magazine, IEEE,* vol. 22, pp. 27-33, 2007.

[6] K. Chen and J. C. Miles, "Its handbook 2004: Recommendations from the world road association (piarc)," 2004.

[7] S. L. Toral, M. R. M. Torres, F. J. Barrero, and M. R. Arahal, "Current paradigms in intelligent transportation systems," *Intelligent Transport Systems, IET,* vol. 4, pp. 201-211, 2010.

[8] H. Xu, "Decentralized Traffic Information System Design Based on Inter-Vehicle Communication," Citeseer, 2006.

[9] L. Wischoff, A. Ebner, H. Rohling, M. Lott, and R. Halfmann, "SOTIS - a self-organizing traffic information system," in *Vehicular Technology Conference, 2003. VTC 2003-Spring. The 57th IEEE Semiannual*, 2003, pp. 2442-2446 vol.4.

[10] P. Raphiphan, W. Pattara-Atikom and P. Prathombutr, "A Survey of Travel Time Estimation Techniques Based on Cellular Probes," in *NSTDA Annual Conference*, Pathumthani, Thailand, 2007.

[11] P. T. Martin, Y. Feng and X. Wang, "Detector Technology Evaluation," Mountain-Plains Consortium, 2003.

[12] Z. Liang, X. Jian-Min and Z. Ling-Xiang, "Arterial speed studies with taxi equipped with global positioning receivers as probe vehicle," in *Wireless*

*Communications, Networking and Mobile Computing, 2005. Proceedings. 2005 International Conference on*, 2005, pp. 1343-1347.

[13] B. Coifman, M. McCord, R. G. Mishalani, M. Iswalt, and Y. Ji, "Roadway traffic monitoring from an unmanned aerial vehicle," *Intelligent Transport Systems, IEE Proceedings,* vol. 153, pp. 11-20, 2006.

[14] L. Stefan, M. Peter, T. Kai-Uwe, C. Dhiraj, P. Bert, and S. Ralf-Peter, "Towards area-wide traffic monitoring-applications derived from probe vehicle data," in *Proceedings of the 8th International Conference on Applications of Advanced Technologies in Transportation Engineering*, Beijing, 2004, pp. 389-394.

[15] S. Thajchayapong, W. Pattara-atikom, N. Chadil, and C. Mitrpant, "Enhanced detection of road traffic congestion areas using cell dwell times," in *Intelligent Transportation Systems Conference, 2006. ITSC '06. IEEE*, 2006, pp. 1084-1089.

[16] L. Yanying and M. McDonald, "Link travel time estimation using single GPS equipped probe vehicle," in *Intelligent Transportation Systems, 2002. Proceedings. The IEEE 5th International Conference on Inteligent Transportation System*, 2002, pp. 932-937.

[17] R. Jean-Gabriel, "Computing travel time-estimates from GSM signalling messages: the STRIP project," 2001.

[18] W. Schneider, "Mobile phones as a basis for traffic state information," in *Intelligent Transportation Systems, 2005. Proceedings. 2005 IEEE*, 2005, pp. 782-784.

[19] R. Geoff, "Mobile Phones as Traffic Probes: Practices, Prospects and Issues," *Transport Reviews,* vol. 26, p. 17, 1/05/2006 2006.

[20] D. J. Dailey and F. W. Cathey, "Virtual speed sensors using transit vehicles as traffic probes," in *Intelligent Transportation Systems, 2002. Proceedings. The IEEE 5th International Conference on*, 2002, pp. 560-565.

[21] K. Choi and Y. Chung, "A Data Fusion Algorithm for Estimating Link Travel Time " *ITS Journal (Intelligent Transportation Systems),* vol. 7, p. 26, 2002-07-01 2002.

[22] Available: http://www.thaieasypass.com/etcsite/index.php?option=com_content&view=section&layout=blog&id=8&Itemid=7&lang=th.

[23] K. Tripruch, "The Application of Advanced Technology for Efficient Traffic Control in Bangkok and Its Vicinities," Master of Sciences, College of Innovation Thammasat University, Bangkok, 2001.

[24] Traffic&Transportation, "Traffic Statistic of Bangkok 2011," Bangkok Metropolitan Administration, Bangkok28/04/2013 2011.

[25] ITS-Groupwork, "Project of ITS Development," Office of Transport and Traffic Policy and Planning, Ministry of Transport, Bangkok2005.

[26] L. Li, C. Long, H. Xiaofei, and H. Jian, "A Traffic Congestion Estimation Approach from Video Using Time-Spatial Imagery," in *Intelligent Networks and Intelligent Systems, 2008. ICINIS '08. First International Conference on*, 2008, pp. 465-469.

[27] P. Zhuang, Y. Shang and B. Hua, "Statistical methods to estimate vehicle count using traffic cameras," *Multidimensional Syst. Signal Process.,* vol. 20, pp. 121-133, 2009.

[28] D. Valerio, A. D'Alconzo, F. Ricciato, and W. Wiedermann, "Exploiting Cellular Networks for Road Traffic Estimation: A Survey and a Research Roadmap," in *Vehicular Technology Conference, 2009. VTC Spring 2009. IEEE 69th*, 2009, pp. 1-5.

[29] W. Pattara-atikom, R. Peachavanish and R. Luckana, "Estimating Road Traffic Congestion using Cell Dwell Time with Simple Threshold and Fuzzy Logic Techniques," in *Intelligent Transportation Systems Conference, 2007. ITSC 2007. IEEE*, 2007, pp. 956-961.

[30] W. Hongsakham, W. Pattara-atikom and R. Peachavanish, "Estimating road traffic congestion from cellular handoff information using cell-based neural networks and K-means clustering," in *Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, 2008. ECTI-CON 2008. 5th International Conference on*, 2008, pp. 13-16.

[31] K. Mandal, A. Sen, A. Chakraborty, S. Roy, S. Batabyal, and S. Bandyopadhyay, "Road traffic congestion monitoring and measurement using active RFID and GSM technology," in *Intelligent Transportation Systems (ITSC), 2011 14th International IEEE Conference on*, 2011, pp. 1375-1379.

[32] A. R. Hevner, "A Three Cycle View of Design Science Research," *Scandinavian journal of information systems,* vol. 19, 2007.

[33] A. R. Hevner, S. T. March, J. Park, and S. Ram, "DESIGN SCIENCE IN INFORMATION SYSTEMS RESEARCH," *MIS Quarterly,* vol. 28, pp. 75-105, 2004.

[34] V. K. Vaishnavi and W. Kuechler Jr, *Design science research methods and patterns: innovating information and communication technology*: CRC Press, 2007.

[35] V. Vaishnavi and W. Kuechler, "Design research in information systems," 2004.

[36] G. D. Abowd, A. K. Dey, P. J. Brown, N. Davies, M. Smith, and P. Steggles, "Towards a Better Understanding of Context and Context-Awareness," in *Proceedings of the 1st international symposium on Handheld and Ubiquitous Computing*, Karlsruhe, Germany, 1999.

[37] W. Huadong, M. Siegel and S. Ablay, "Sensor fusion for context understanding," in *Instrumentation and Measurement Technology Conference, 2002. IMTC/2002. Proceedings of the 19th IEEE*, 2002, pp. 13-17 vol.1.

[38] S. K. Endarnoto, S. Pradipta, A. S. Nugroho, and J. Purnama, "Traffic Condition Information Extraction & Visualization from Social Media Twitter for Android Mobile Application," in *Electrical Engineering and Informatics (ICEEI), 2011 International Conference on*, 2011, pp. 1-4.

[39] R. Kosala, E. Adi and Steven, "Harvesting Real Time Traffic Information from Twitter," *Procedia Engineering,* vol. 50, pp. 1-11, 2012.

[40] W. M. Trochim, "Likert scaling," *Research methods knowledge base,* vol. 2, 2006.

[41] L. Figueiredo, I. Jesus, J. A. T. Machado, J. R. Ferreira, and J. L. Martins de Carvalho, "Towards the development of intelligent transportation systems," in *Intelligent Transportation Systems, 2001. Proceedings. 2001 IEEE*, 2001, pp. 1206-1211.

[42] I. Masaki, "Machine-vision systems for intelligent transportation systems," *Intelligent Systems and their Applications, IEEE,* vol. 13, pp. 24-31, 1998.

[43] "Advanced Vehicle and Highway Technologies, TRB Special Report 232," Transportation Research Board, Nat'l Research Council, Washington D.C.1991.

[44] V. Graefe and K.-D. Kuhnert, "Vision-based autonomous road vehicles," in *Vision-based vehicle guidance*, M. Ichiro, Ed., ed: Springer-Verlag New York, Inc., 1992, pp. 1-29.

[45] B. Ulmer, "VITA II-active collision avoidance in real traffic," in *Intelligent Vehicles '94 Symposium, Proceedings of the*, 1994, pp. 1-6.

[46] M. Satyanarayanan, "Pervasive computing: vision and challenges," *Personal Communications, IEEE,* vol. 8, pp. 10-17, 2001.

[47] L. A. Klein, *Sensor Technologies and Data Requirements for ITS Applications*: Artech House Publishers, 2001.

[48] J. L. Ygnace and C. Drane, "Cellular telecommunication and transportation convergence: a case study of a research conducted in California and in France on cellular positioning techniques and transportation issues," in *Intelligent*

*Transportation Systems, 2001. Proceedings. 2001 IEEE*, Oakland, 2001, pp. 16-22.

[49] S. Riter and J. McCoy, "Automatic vehicle location-An overview," *Vehicular Technology, IEEE Transactions on,* vol. 26, pp. 7-11, 1977.

[50] N. Ayland and P. Davies, "Automatic vehicle identification for heavy vehicle monitoring," in *Road Traffic Monitoring, 1989., Second International Conference on*, 1989, pp. 152-155.

[51] L. Yanping, G. Dudu, S. Mingliang, and S. Qi, "A new idea of identification of road traffic conditions," in *Computer Application and System Modeling (ICCASM), 2010 International Conference on*, 2010, pp. V14-638-V14-642.

[52] W. C. M. Hsiao and S. K. J. Chang, "Segment based traffic information estimation method using cellular network data," in *Intelligent Transportation Systems, 2005. Proceedings. 2005 IEEE*, 2005, pp. 142-147.

[53] F. W. Cathey and D. J. Dailey, "Implementation of traffic management using transit probes," in *Intelligent Transportation Systems, 2005. Proceedings. 2005 IEEE*, 2005, pp. 243-247.

[54] P. Amir, Z. Arkady and L. Seng, "A Unifying Model for Representing and Reasoning About Context under Uncertainty," in *11th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU)*, France, 2006.

[55] M. Aftabuzzaman, "Understanding the road traffic congestion relief impacts of public transport," 2011.

[56] M. J. Rothenberg, "Urban Congestion in the United States: What Does the Future Hold?," 1985.

[57] G. Weisbrod, D. Vary and G. Treyz, "Measuring economic costs of urban traffic congestion to business," *Transportation Research Record: Journal of the Transportation Research Board,* vol. 1839, pp. 98-106, 2003.

[58] Y.-J. Lee and V. R. Vuchic, "Transit network design with variable demand," *Journal of Transportation Engineering,* vol. 131, pp. 1-10, 2005.

[59] Z. Fei and L. Liangyou, "An Optimized Video-Based Traffic Congestion Monitoring System," in *Knowledge Discovery and Data Mining, 2010. WKDD '10. Third International Conference on*, 2010, pp. 150-153.

[60] P. Widhalm, M. Piff, N. Brandle, H. Koller, and M. Reinthaler, "Robust road link speed estimates for sparse or missing probe vehicle data," in *Intelligent Transportation Systems (ITSC), 2012 15th International IEEE Conference on*, 2012, pp. 1693-1697.

[61] R. Bauza, J. Gozalvez and J. Sanchez-Soriano, "Road traffic congestion detection through cooperative Vehicle-to-Vehicle communications," in *Local Computer Networks (LCN), 2010 IEEE 35th Conference on*, 2010, pp. 606-612.

[62] B. Wang, L. Xu and W. Zhang, "Cloud Theory Based Traffic State Identification for Urban Road Network," in *Intelligent Computation Technology and Automation (ICICTA), 2010 International Conference on*, 2010, pp. 946-949.

[63] P. Kachroo, K. Ozbay and A. G. Hobeika, "Real-time travel time estimation using macroscopic traffic flow models," in *Intelligent Transportation Systems, 2001. Proceedings. 2001 IEEE*, 2001, pp. 132-137.

[64] A. Phan and F. Ferrie, "Obtaining Dense Road Speed Estimates from Sparse GPS Measurements," in *Intelligent Transportation Systems, 2008. ITSC 2008. 11th International IEEE Conference on*, 2008, pp. 157-162.

[65] Z. Sun, M. Guo, W. Liu, J. Feng, and J. Hu, "Multisource Traffic Data Fusion with Entropy Based Method," in *Artificial Intelligence and Computational Intelligence, 2009. AICI '09. International Conference on*, 2009, pp. 506-509.

[66] X. Ma and H. Koutsopoulos, "Estimation of the automatic vehicle identification based spatial travel time information collected in Stockholm," *Intelligent Transport Systems, IET,* vol. 4, pp. 298-306.

[67] S. Tao, V. Manolopoulos, S. Rodriguez, and A. Rusu, "Real-Time Urban Traffic State Estimation with A-GPS Mobile Phones as Probes," *Journal of Transportation Technologies,* vol. 2, p. 22, 2012.

[68] H. van Lint, O. Miete, H. Taale, and S. Hoogendoorn, "Systematic Framework for Assessing Traffic Measures and Policies on Reliability of Traffic Operations and Travel Time," *Transportation Research Record: Journal of the Transportation Research Board,* vol. 2302, pp. 92-101, 2012.

[69] S. Shang, H. Lu, T. B. Pedersen, and X. Xie, "Modeling of traffic-aware travel time in spatial networks," in *Mobile Data Management (MDM), 2013 IEEE 14th International Conference on*, 2013, pp. 247-250.

[70] H. Tan, G. Feng, J. Feng, W. Wang, Y.-J. Zhang, and F. Li, "A tensor-based method for missing traffic data completion," *Transportation Research Part C: Emerging Technologies,* vol. 28, pp. 15-27, 2013.

[71] Z. Yang and L. Yuncai, "Missing traffic flow data prediction using least squares support vector machines in urban arterial streets," in *Computational Intelligence and Data Mining, 2009. CIDM '09. IEEE Symposium on*, 2009, pp. 76-83.

[72] J. Reddy, S. Finger, S. Konda, and E. Subrahmanian, "Design as Building and Reusing Artifact Theories: Understanding and Supporting Growth of Design Knowledge," in *The Design Productivity Debate*, A. B. Duffy, Ed., ed: Springer London, 1998, pp. 268-290.

[73] K. Ron and P. Foster, "Special issue on applications of machine learning and the knowledge discovery process," *Journal of Machine Learning,* vol. 30, pp. 271-274, 1998.

[74] Z. Liu, S. Sharma and S. Datla, "Imputation of Missing Traffic Data during Holiday Periods," *Transportation Planning and Technology,* vol. 31, pp. 525-544, 2008/10/01 2008.

[75] M. Weiser and J. S. Brown, "The coming age of calm technolgy," in *Beyond calculation*, ed: Copernicus, 1997, pp. 75-85.

[76] F. X. company. *Ubiquitous Area*. Available: http://www.fujixerox.com/eng/company/technology/ubique.html.

[77] L. Eunyoung, K. Ryu and I. Paik, "A Concept for Ubiquitous Transportation Systems and Related Development Methodology," in *Intelligent Transportation Systems, 2008. ITSC 2008. 11th International IEEE Conference on*, 2008, pp. 37-42.

[78] H. Gellersen, A. Schmidt and M. Beigl, "Multi-Sensor Context-Awareness in Mobile Devices and Smart Artifacts," *Mobile Networks and Applications,* vol. 7, pp. 341-351, 2002/10/01 2002.

[79] A. Padovitz, "Context management and reasoning about situations in pervasive computing," Monash University, 2006.

[80] B. Schilit, N. Adams and R. Want, "Context-aware computing applications," in *Mobile Computing Systems and Applications, 1994. Proceedings., Workshop on*, 1994, pp. 85-90.

[81] A. Achilleos, K. Yang and N. Georgalas, "Context modelling and a context-aware framework for pervasive service creation: A model-driven approach," *Pervasive and Mobile Computing,* vol. 6, pp. 281-296, 2010.

[82] J. Reid, R. Hull, B. Clayton, G. Porter, and P. Stenton, "Priming, sense-making and help: Analysis of player behaviour in an immersive theatrical experience," *Pervasive Mob. Comput.,* vol. 6, pp. 499-511.

[83] C. Dominguez, M. Vidulich, E. Vogel, and G. McMillan, "{Can SA be defined?}," in *{Situation Awareness: Papers and Annotated Bibliography. Interim Report No. AL/CF-TR-1994-0085}*, ed, 1994, pp. 5-15.

[84] S. Nava-Muñoz and A. L. Morán, "CANoE: A context-aware notification model to support the care of older adults in a nursing home," *Sensors,* vol. 12, pp. 11477-11504, 2012.

[85] J. Floch, C. Frà, R. Fricke, K. Geihs, M. Wagner, J. Lorenzo, E. Soladana, S. Mehlhase, N. Paspallis, and H. Rahnama, "Playing MUSIC—building context-aware and self-adaptive mobile applications," *Software: Practice and Experience,* vol. 43, pp. 359-388, 2013.

[86] K. Geihs, R. Reichle, M. Wagner, and M. U. Khan, "Modeling of context-aware self-adaptive applications in ubiquitous and service-oriented environments," in *Software engineering for self-adaptive systems*, ed: Springer, 2009, pp. 146-163.

[87] U. Paul, A. P. Subramanian, M. M. Buddhikot, and S. R. Das, "Understanding traffic dynamics in cellular data networks," in *INFOCOM, 2011 Proceedings IEEE*, 2011, pp. 882-890.

[88] M. Proebster, M. Kaschub, T. Werthmann, and S. Valentin, "Context-aware resource allocation for cellular wireless networks," *EURASIP Journal on Wireless Communications and Networking,* vol. 2012, pp. 1-19, 2012/07/12 2012.

[89] D. T. Larose, *Discovering Knowledge in Data : An Introduction to Data Mining*, 1 ed. Hoboken: Wiley-Interscience 2005.

[90] F. Provost and T. Fawcett, *Data Science for Business : What You Need to Know about Data Mining and Data-analytic Thinking*. Sebastopol, CA, USA: O'Reilly Media, 2013.

[91] C. Shearer, "The CRISP-DM model: the new blueprint for data mining," *Journal of data warehousing,* vol. 5, pp. 13-22, 2000.

[92] M. Sharma, "Data Mining: A Literature Survey."

[93] E. Alpaydin, *Introduction to machine learning*: MIT press, 2004.

[94] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition*. USA: Morgan Kaufmann 2005.

[95] M. Kantardzic, *Data mining: concepts, models, methods, and algorithms*, 2nd ed.: John Wiley & Sons, 2011.

[96] V. Cherkassky and F. M. Mulier, *Learning from data: concepts, theory, and methods*: John Wiley & Sons, 2007.

[97] R. Caruana and A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms," in *Proceedings of the 23rd international conference on Machine learning*, Pittsburgh, Pennsylvania, 2006, pp. 161-168.

[98] A. M. Andrew, "Backpropagation," *Kybernetes,* vol. 30, pp. 1110-1117, 2001.

[99] D. E. Rumelhart, G. E. Hinton and R. J. Williams, "Learning internal representations by error propagation," DTIC Document1985.

[100] A. Aamodt and E. Plaza, "Case-based reasoning: Foundational issues, methodological variations, and system approaches," *AI communications,* vol. 7, pp. 39-59, 1994.

[101] T. M. Mitchell, *Machine Learning*: McGraw-Hill, Inc., 1997.

[102] J. R. Quinlan, "Induction of decision trees," *Machine learning,* vol. 1, pp. 81-106, 1986.

[103] J. R. Quinlan, *C4.5: Programs for Machine Learning*: Morgan Kaufmann, 1993.

[104] J. R. Quinlan, "Simplifying decision trees," *International Journal of Man-Machine Studies,* vol. 27, pp. 221-234, 1987.

[105] W. W. Cohen, "Fast Effective Rule Induction " in *In Proceedings of the Twelfth International Conference on Machine Learning*, 1995.

[106] S.-B. Park, "Combining rule-based learning and memory-based learning for automatic word spacing in simple message service," *Appl. Soft Comput.,* vol. 6, pp. 406-416, 2006.

[107] D. Barber, *Bayesian Reasoning and Machine Learning*: Cambridge University Press, 2012.

[108] S.-B. Kim, K.-S. Han, H.-C. Rim, and S. H. Myaeng, "Some Effective Techniques for Naive Bayes Text Classification," *IEEE Trans. on Knowl. and Data Eng.,* vol. 18, pp. 1457-1466, 2006.

[109] S. Z. Selim and M. A. Ismail, "K-means-type algorithms: a generalized convergence theorem and characterization of local optimality," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* pp. 81-87, 1984.

[110] R. G. D'Andrade, "U-statistic hierarchical clustering," *Psychometrika,* vol. 43, pp. 59-67, 1978.

[111] Z. Ghahramani, "An introduction to hidden Markov models and Bayesian networks," *International Journal of Pattern Recognition and Artificial Intelligence,* vol. 15, pp. 9-42, 2001.

[112] R. J. Elliott, L. Aggoun and J. B. Moore, *Hidden Markov Models*: Springer, 1994.

[113] D. Arnott, "Cognitive biases and decision support systems development: a design science approach," *Information Systems Journal,* vol. 16, pp. 55-78, 2006.

[114] H. A. Simon, *The sciences of the artificial* vol. 136: MIT press, 1969.

[115] S. T. March and G. F. Smith, "Design and natural science research on information technology," *Decision support systems,* vol. 15, pp. 251-266, 1995.

[116] H. Takeda, P. Veerkamp and H. Yoshikawa, "Modeling design process," *AI magazine,* vol. 11, p. 37, 1990.

[117] P.-N. Tan, M. Steinbach and V. Kumar, *Introduction to Data Mining* USA: Addison-Wesley, 2005.

[118] A. Ranganathan, J. Al-Muhtadi and R. H. Campbell, "Reasoning about uncertain contexts in pervasive computing environments," *Pervasive Computing, IEEE,* vol. 3, pp. 62-70, 2004.

[119] G. Chen and D. Kotz, "A Survey of Context-Aware Mobile Computing Research," Dartmouth College, Hanover2000.

[120] S. M. Kothuri, K. Tufte, A. Soyoung, and R. L. Bertini, "Development of an ITS data archive application for improving freeway travel time estimation," in *Intelligent Transportation Systems Conference, 2006. ITSC '06. IEEE*, 2006, pp. 1263-1268.

[121] P. Raphiphan, A. Zaslavsky, P. Prathombutr, and P. Meesad, "Overcoming uncertainty of roadside sensors with smart adaptive traffic congestion analysis system," in *Intelligent Vehicles Symposium, 2009 IEEE*, 2009, pp. 1045-1050.

[122] P. Raphiphan, A. Zaslavsky, P. Prathombutr, and P. Meesad, "Context Aware Traffic Congestion Estimation to Compensate Intermittently Available Mobile Sensors," in *Mobile Data Management: Systems, Services and Middleware, 2009. MDM '09. Tenth International Conference on*, 2009, pp. 405-410.

[123] P. Raphiphan, P. Prathombutr, A. Zaslavsky, and P. Meesad, "Real time traffic congestion degree computation for minor sensorless roads using cost efficient context reasoning," in *Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference on*, 2010, pp. 1741-1746.

[124] Y. Zhu, J. Wang and H. Lu, "A Study on Urban Traffic Congestion Dynamic Predict Method Based on Advanced Fuzzy Clustering Model," in *Computational Intelligence and Security, 2008. CIS '08. International Conference on*, 2008, pp. 96-100.

[125] H. Bingwei, X. Shuangjiu, L. Lu, and W. Zhijing, "A new method for filling missing values by gray relational analysis," in *Artificial Intelligence,*

*Management Science and Electronic Commerce (AIMSEC), 2011 2nd International Conference on*, 2011, pp. 2721-2724.

[126] G. Madhu and T. V. Rajinikanth, "A novel index measure imputation algorithm for missing data values: A machine learning approach," in *Computational Intelligence & Computing Research (ICCIC), 2012 IEEE International Conference on*, 2012, pp. 1-7.

[127] *Groovy Home*. Available: http://groovy.codehaus.org.

[128] V. Vaithiyanathan, K. Rajeswari, K. Tajane, and R. Pitale, "Comparison of Different Classification Techniques Using Different Datasets," 1963.

[129] D. Soria, J. M. Garibaldi, F. Ambrogi, E. M. Biganzoli, and I. O. Ellis, "A 'non-parametric' version of the naive Bayes classifier," *Knowledge-Based Systems,* vol. 24, pp. 775-784, 2011.

[130] "iTIC Bangkok Location Table Version 2," ed. Bangkok: The Intelligent Traffic Information Center Foundation.

[131] *Pressure Hose Vehicle Counter*. Available: http://www.vehicle-counters.com/vehicle-counter-hose-sensors.htm.

[132] *Waze*. Available: http://world.waze.com/.

[133] P. Posawang, S. Phosaard, W. Pattara-Atikom, and W. Polnigongit, "Perception-based Road Traffic Congestion Classification using Neural Networks," in *International Conference on World Congress of Engineering (WCE 2009), London, England*, 2009.

[134] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Technique*, 2 ed. USA: Morgan Kaufmann, 2005.

[135] M. Saar-Tsechansky and F. Provost, "Handling missing values when applying classification models," 2007.

[136] J. W. Grzymala-Busse and M. Hu, "A comparison of several approaches to missing attribute values in data mining," in *Rough sets and current trends in computing*, 2001, pp. 378-385.

[137] D. Frippiat and N. Marquis, "Web surveys in the social sciences: An overview," *Population (english edition),* vol. 65, pp. 285-311, 2010.

[138] M. Couper, *Designing effective Web surveys*. Cambridge ; New York: Cambridge University Press, 2008.

[139] D. A. Dillman, *Mail and internet surveys: The tailored design method* vol. 2: Wiley New York, 2000.

[140] K. B. Wright, "Researching Internet-based populations: Advantages and disadvantages of online survey research, online questionnaire authoring software packages, and web survey services," *Journal of Computer-Mediated Communication,* vol. 10, pp. 00-00, 2005.

[141] M. Oppermann, "E-mail surveys-potentials and pitfalls," *Marketing Research,* vol. 7, pp. 29-33, 1995.

[142] M. P. Couper, "Review: Web surveys: A review of issues and approaches," *The Public Opinion Quarterly,* vol. 64, pp. 464-494, 2000.

[143] L. Frazer, *Questionnaire design and administration : a practical guide*. Brisbane: Brisbane : John Wiley & Sons Australia, 2000.

[144] P. D. o. D. o. L. T. o. T. Transports Statistics Sub-Division, "The Number of Driving Licences and Transport Personnel Licences Classified by Type As of December 2012," Department of Land Transport of Thailand, Bangkok2012.

[145] R. V. Krejcie and D. W. Morgan, "Determining sample size for research activities," *Educational and psychological measurement,* vol. 30, pp. 607-610, 1970.

[146] T. Yamane, *Statistics: an introductory analysis*. New York: New York : Harper & Row, 1973.

[147] P. F. Rad, "Project Estimating and Cost Management," ed: Management Concepts, Inc.

[148] O. Zwikael and S. Globerson, "Evaluating the quality of project planning: a model and field results," *International Journal of Production Research,* vol. 42, pp. 1545-1556, 2004.

[149] *Thailand Yearbook of Telecomunications indicators : 2012-2013*. Bangkok: Office of National Broadcasting and Telecommunications Commission (NBTC), 2013.

[150] "Report of Telecomunication Market in the 3rd quater of year 2013," Office of National Broadcasting and Telecommunications Commission (NBTC) Bangkok2013.

[151] I. E. Allen and C. A. Seaman, "Likert scales and data analyses," *Quality Progress,* vol. 40, pp. 64-65, 2007.

[152] *Traffy*. Available: www.twitter.com/traffy.

[153] C. C. Aggarwal, *Social network data analytics*: New York : Springer 2011.

[154] P. J. Denning, "A new social contract for research," *Commun. ACM,* vol. 40, pp. 132-134, 1997.

[155] D. Tsichritzis, "The dynamics of innovation," in *Beyond calculation*, ed: Springer, 1997, pp. 259-265.

[156] V. R. Basili, "The role of experimentation in software engineering: past, current, and future," in *Proceedings of the 18th international conference on Software engineering*, 1996, pp. 442-449.

[157] M. V. Zelkowitz and D. R. Wallace, "Experimental models for validating technology," *Computer*, vol. 31, pp. 23-31, 1998.

# Appendix A : Experiment Results from the Single Model Approach When Applying only Context Attributes in *RS*

For this experiment in Evaluation II involving the single model approach, we build a model from the context attributes in *RS*, evaluate each model ten times and then calculate the average accuracy and average learning and model building time. As in Section 5.3 in Chapter 5, the reason for repeating the task ten times and calculating the average is that the data for testing differs in each run. This is because we randomly insert NULL values to simulate the missing data. The average accuracy and average model building time computed for each road segment is presented in Table A-1.

**Table A-1: Average accuracy and model building time (single model approach) when applying context attributes from *RS***

| Road Link | Accuracy (%) | | Model Building Time (μs) | |
|---|---|---|---|---|
| | Mean | S.D. | Mean | S.D. |
| 1206 | 93.75 | 0.08 | 7049 | 1015 |
| 2613 | 87.23 | 0.13 | 9418 | 1449 |
| 2718 | 98.97 | 0.01 | 7529 | 1835 |

After building an inference model from the context attributes in *RS*, this model is used to infer the missing sensory traffic data of an observed road segment with different rates of missing data (20%, 40%, 60%, 80% and 100% respectively). We ran the experiment ten times and show only the average of those ten runs in this section. Table A-2 and the graph in Figure A-1 show the average time used to infer the inferred traffic congestion degree when the sensory traffic data of a particular road segment is missing. Table A-3 shows the average accuracy and SD at different

missing data rates for each road segment. We also show the results of the active and non-active periods in Table A-4 and Table A-5 respectively.

**Table A-2: Inferring time (μs) for specific missing data rates (%) for the single model approach when applying *RS***

| Road Link | Inferring time (μs) at specific data missing rates (%) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 20% | | 40% | | 60% | | 80% | | 100% | |
| | Mean | S.D. | Mean | S.D. | Mean | S.D. | Mean | S.D. | Mean | S.D. |
| 1206 | 1329 | 6 | 1331 | 5 | 1336 | 6 | 1330 | 8 | 1320 | 2 |
| 2613 | 1903 | 39 | 1901 | 9 | 1908 | 6 | 1909 | 11 | 1898 | 10 |
| 2718 | 1036 | 3 | 1041 | 4 | 1040 | 3 | 1031 | 5 | 1025 | 3 |

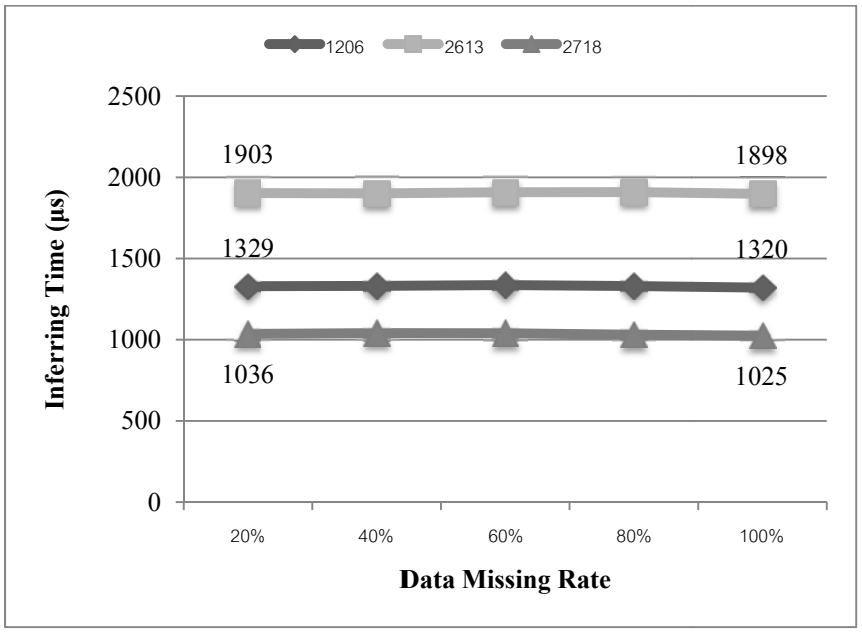



**Figure A-1: Inferring time (μs) at specific missing data rates (%) for the single model approach when applying only *RS***

**Table A-3: Accuracy (%) at specific missing data rates (%) for the single model approach when applying only *RS***

| Road Link | Accuracy (%) at specific missing data rates (%) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 20% | | 40% | | 60% | | 80% | | 100% | |
| | Mean | S.D. | Mean | S.D. | Mean | S.D. | Mean | S.D. | Mean | S.D. |
| 1206 | 88.07 | 0.44 | 88.30 | 0.14 | 88.27 | 0.15 | 88.36 | 0.09 | 88.31 | 0.00 |
| 2613 | 80.25 | 0.26 | 79.60 | 0.21 | 78.53 | 0.17 | 77.14 | 0.13 | 75.46 | 0.00 |
| 2718 | 96.58 | 0.11 | 94.13 | 0.14 | 91.70 | 0.17 | 89.27 | 0.07 | 86.83 | 0.00 |

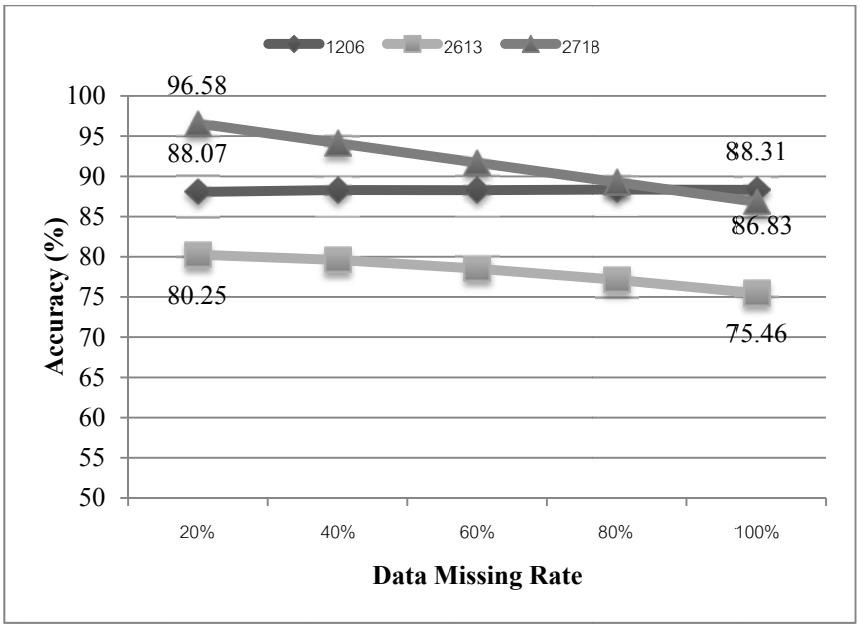



**Figure A-2: Accuracy (%) at specific missing data rates (%) for the single model approach when applying only *RS***

**Table A-4: Accuracy (%) of the *active period* at specific missing data rates (%) for the single model approach when applying only *RS***

| Road Link | Active period accuracy (%) at specific data missing rate (%) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 20% | | 40% | | 60% | | 80% | | 100% | |
| | Mean | S.D. | Mean | S.D. | Mean | S.D. | Mean | S.D. | Mean | S.D. |
| 1206 | 80.45 | 0.71 | 80.81 | 0.27 | 80.86 | 0.27 | 81.01 | 0.15 | 81.00 | 0.00 |
| 2613 | 62.93 | 0.53 | 61.35 | 0.45 | 58.71 | 0.27 | 55.40 | 0.22 | 51.33 | 0.00 |
| 2718 | 93.49 | 0.27 | 88.75 | 0.29 | 84.08 | 0.30 | 79.43 | 0.15 | 74.75 | 0.00 |

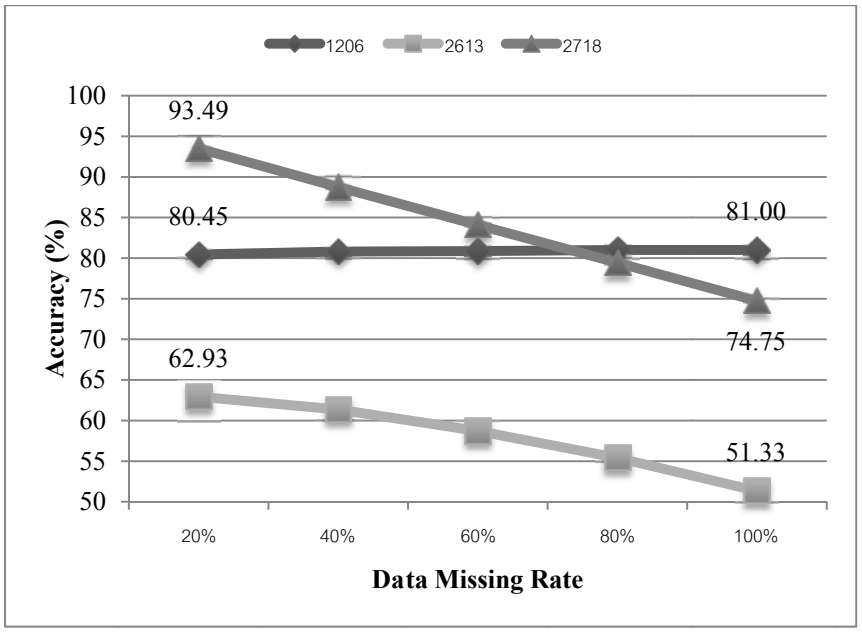



**Figure A-3: Accuracy (%) of the *active period* at specific missing data rates (%) for the single model approach when applying only *RS***

**Table A-5: Accuracy (%) of the *non-active period* at specific missing data rates (%) for the single model approach when applying only *RS***

| Road Link | Non-active period accuracy (%) at specific missing data rates (%) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 20% | | 40% | | 60% | | 80% | | 100% | |
| | Mean | S.D. | Mean | S.D. | Mean | S.D. | Mean | S.D. | Mean | S.D. |
| 1206 | 95.55 | 0.35 | 95.67 | 0.17 | 95.56 | 0.12 | 95.59 | 0.06 | 95.50 | 0.00 |
| 2613 | 97.20 | 0.17 | 97.56 | 0.12 | 98.04 | 0.08 | 98.56 | 0.05 | 99.20 | 0.00 |
| 2718 | 99.64 | 0.10 | 99.41 | 0.08 | 99.21 | 0.05 | 98.95 | 0.04 | 98.73 | 0.00 |

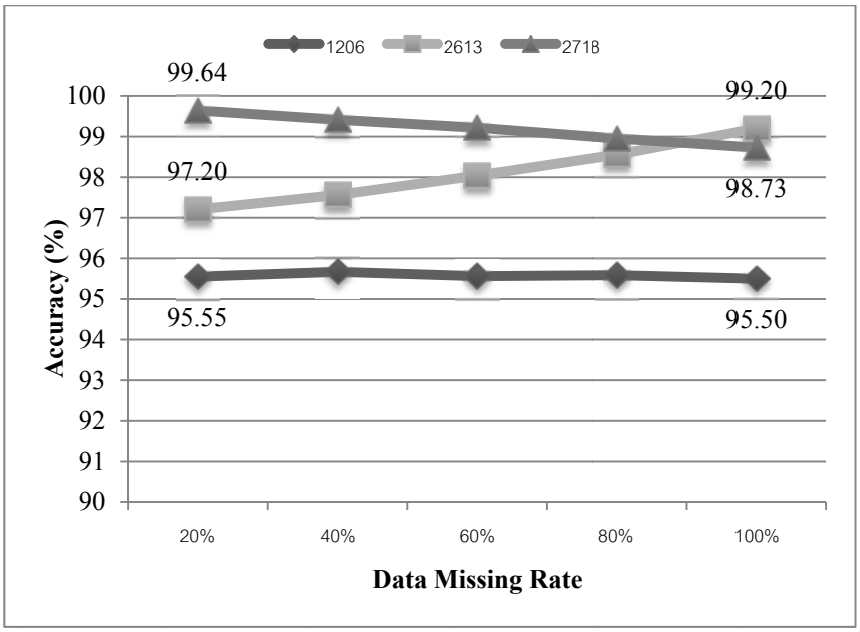



**Figure A-4: Accuracy (%) of the *non-active period* at specific missing data rates (%) for the single model approach when applying only *RS***

Discussion

In Table A-1, the accuracy in model evaluation when using context attributes from *RS* are similar to those in Table 5-10 when all context attributes are used. The model building time differs very little between the two sets of results.

The inferring time shown in Table A-2 (when only context attributes from *RS* are used) is less than the inferring time shown in Table 5-11 (when all context attributes

are used). This is because of the smaller number of context attributes, which in turn reduces the size of the decision tree and thus minimizes the computation time.

The results from this experiment show that the overall accuracy, as well as the accuracy in the active and non-active periods (illustrated in Figure A-2, Figure A-3 and Figure A-4) are quite similar to the results of the experiment reported in Figure 5-22, Figure 5-23 and Figure 5-24 in Chapter 5. Similar accuracy levels are achieved whether applying all context attributes or whether applying only context attributes from *RS* in the single model approach.

# Appendix B : Implementation of the Proposed Framework (Source Code with Explanation)

```
/*
 * Copyright (C) 2014 @panraphee
 *
 */

//Configure Dependency
// @Grapes([
//     @Grab(group='nz.ac.waikato.cms.weka', module='weka-stable', version='3.6.11'),
//     @Grab(group='joda-time', module='joda-time', version='1.6.1')
// ])


//import weka class
import weka.classifiers.Classifier
import weka.classifiers.Evaluation
import weka.classifiers.meta.FilteredClassifier
import weka.classifiers.misc.SerializedClassifier
import weka.classifiers.rules.JRip
import weka.classifiers.trees.J48
import weka.core.Attribute
import weka.core.Debug
import weka.core.Instance
import weka.core.Instances
import weka.core.converters.ConverterUtils.DataSource
import weka.filters.unsupervised.attribute.Remove

//import joda for time calculation purpose
import org.joda.time.DateTime
import org.joda.time.format.DateTimeFormat
import org.joda.time.format.ISODateTimeFormat


/*
 *
 * @author @panraphee
 */
```

```
class RoadLink {

//Attributes
    //settings, assign at new. var with default value can be omitted
    def name
    def algorithm = 'J48'
    def numberOfRecordToLearn = 0
    def learningSched = 0
    def missRate = 0
    def mode = 3
    def seed = 1

    //context data member
    DateTime logTime
    def day
    def time
    def rain_mm
    def rain
    def school
    def connectedRoadLink1
    def connectedRoadLink2
    def roadLink

    // internal use data
    def recordNo
    def date
    def learnNo = 0
    def roadLinkInferred
    def roadLinkActual
    def connectedRoadLink1Name
    def connectedRoadLink2Name
    boolean NeedToReLearn
    Instances structure

    def recordWithActualVal = 0
    def correct = 0
    def currentCorrect = []
    def accuracy = ''
    def currentAccuracy = ''

    def reportCorrect = 0
    def recordWithMissingVal = 0
    def reportAccuracy
    def reportCorrectA = 0
    def recordWithMissingValA = 0
    def reportAccuracyA
    def reportCorrectN = 0
    def recordWithMissingValN = 0
    def reportAccuracyN



    def totalInferTime = 0

    def inferLog = "No,Log Time, Model, Traffic, Inferred, Point, Correct, " +
                        "RecordWithActual, Accuracy, CurrentAccuracy\n"
    def reportLog = "Log Time, Traffic, Inferred, Point, Actual, Correct, " +
                        "Missing, Accuracy\n"
    def historyLog = ''
```

```
//Methods

//Main Method
    def process() {

        init()

        def dataFile = new File("./ex_data/${name}sensor.and.context.csv")


        dataFile.eachLine {line, number ->
                //skip first line
                if(number == 1) return
                //simulate waiting time to get new data
                //sleep 500

                getData(line, number)
                randomMissing(missRate)
                infer()
                reportTraffic()
                saveHistory()
                if(NeedToReLearn) learn(algorithm, numberOfRecordToLearn)
        }

        summarize()

    }

//Sub Methods
    def init () {


        reset()

        //copy data, keep the original clean
        def src = new File("./ex_data/${name}history.arff")
        def dest = new File("./data/${name}${algorithm}/${name}history.arff")
        dest.write(src.text)


        //learn from history data
        learn(algorithm, numberOfRecordToLearn)

    }
```

```
    def reset() {

        new File("./data/${name}${algorithm}/models").deleteDir()
        new File("./data/${name}${algorithm}").deleteDir()

        new File("./data/${name}${algorithm}").mkdir()
        new File("./data/${name}${algorithm}/models/").mkdir()

    }
```

```
def learn(algo, numberOfRecordToLearn){

    learnNo++

    //create instances
    def s = new File("./data/${name}${algorithm}/${name}history.arff")
    DataSource source = new DataSource(s.path)
    Instances trainInstances = source.getDataSet()

    //get empty instances with full arff structure for later ref
    structure = trainInstances.stringFreeStructure()

    //get connected roadlink name to use later
    connectedRoadLink1Name = trainInstances.attribute(6).name()
    connectedRoadLink2Name = trainInstances.attribute(7).name()



    //remove instance in reverse order, prevent order shift
    // if 0, learn all
    if(numberOfRecordToLearn > 0) {

            int deleteUntilRecord = (trainInstances.numInstances() - 1) -
                                    numberOfRecordToLearn
            for(int i = deleteUntilRecord; i >= 0; i--) {
                trainInstances.delete(i)
            }

    }


    int cIdx=trainInstances.numAttributes() - 1
    trainInstances.setClassIndex(cIdx)

    J48 j48 = new J48()
    j48.setUnpruned(true)

    JRip jrip = new JRip()

    def defaultRemoveCTX
    def toRemoveCTX

    switch (mode) {
        case 0:
            defaultRemoveCTX = '1,4'
            toRemoveCTX = []
            break
        case 1:
            defaultRemoveCTX = '1,4'
            toRemoveCTX = [2, 3, 5, 6, 7, 8]
            break
        case 2:
            defaultRemoveCTX = '1,4,5,6'
            toRemoveCTX = []
            break
        default:
            defaultRemoveCTX = '1,4,5,6'
            toRemoveCTX = [7, 8]
    }
```

```
def missingCombinations = toRemoveCTX.subsequences()*.combinations().
                                                        inject( [] ) { list, set ->
list.addAll( set )
list
}*.join(",").sort { it.length() }

missingCombinations += ''
if (mode == 1) {missingCombinations -= toRemoveCTX.join(",")}

println "\nBuilding $name ${missingCombinations.size} $algorithm " +
            "models with ${trainInstances.numInstances()} instances"
println ' . '* (missingCombinations.size/3)

def totalBuildTime = 0
def totalEvalTime = 0
def totalTime = 0

missingCombinations.eachWithIndex { item, index ->
    //make filter to remove missing attribute
    Remove rm = new Remove()
    String missingVal = item.toString() != '' ? ',' + item.toString():''
    rm.setAttributeIndices(defaultRemoveCTX + missingVal)

    //make filtered classifier
    FilteredClassifier fc = new FilteredClassifier()
    fc.setFilter(rm)
    if(algo == "J48") {

            fc.setClassifier(j48)

    } else {

            fc.setClassifier(jrip)
    }

    //build model, evaluate and record time in microsec
    def buildStart= System.nanoTime()

        fc.buildClassifier(trainInstances)
        if(index%3 == 0) print ' . '

    def buildTime = (System.nanoTime() - buildStart)/1000
    totalBuildTime += buildTime


    //do 10 fold cross validation evaluate
    def evalStart= System.nanoTime()

        Evaluation eval = new Evaluation(trainInstances)
        Random rand = new Random(seed)  // using seed = 1
        int folds = 10
        eval.crossValidateModel(fc, trainInstances, folds, rand)

    def evalTime = (System.nanoTime() - evalStart)/1000
    totalEvalTime += evalTime

    totalTime = totalTime + buildTime + evalTime



    //save model
    def isWritingOk = Debug.saveToFile("./data/${name}${algorithm}" +
                        "/models/${name}.${algorithm}-"+
                        missingVal.replaceAll(',','')+".model", fc)
```

```
        //write tree and statistics to file
        def modelTextName = "${name}.${algorithm}-"+
                        missingVal.replaceAll(',', '')+
                        ".build$learnNo"

        def modelText = new File("./data/${name}${algorithm}/models" +
                                 "/${name}.${algorithm}-"+
                                  missingVal.replaceAll(',', '')+
                                  ".build$learnNo" + ".txt")

        modelText.write("Time taken to build model:     ${buildTime} micro sec\n" +
                        "Time taken to evaluate model:  ${evalTime} micro sec\n" +
                        "Total time used:               ${buildTime + evalTime}" +
                                                                  " micro sec\n\n" +
                        "\n----Evaluation----\n" + eval.toSummaryString() +
                        "\n" + eval.toClassDetailsString() +
                        "\n" + eval.toMatrixString() +
                        "\n\n----Model Description----\n\n" + fc.toString())

        //write build log
        def modelEvaluationAccuracy =  eval.pctCorrect()

        def buildLog = new File("./final/buildLog.csv")
        if (mode == 0 | mode == 2) {
            buildLog << "$modelEvaluationAccuracy, $buildTime, " +
                                 "$evalTime, ${buildTime + evalTime},"
        } else {
           buildLog << "$missRate, $name, $learnNo, ${index+1}, $modelTextName, " +
                               "${missingVal.replaceAll(',', '')}, $modelEvaluationAccuracy, " +
                                 "$buildTime, $evalTime, ${buildTime + evalTime}\n"
        }

    }

    println "\nTotal Building and Evalating Time: ${totalTime} micro sec\n"


    def f = new File("./data/${name}${algorithm}/${name}learnLog.csv")

    //Log: No, Link, Algorithm, Models, Records, Total build time,
     //Total eval time, Total time in micro sec

    def learnLog =  "$learnNo, $name, $algorithm, ${missingCombinations.size},," +
                            "${trainInstances.numInstances()}, "+
                      "${totalBuildTime}, ${totalEvalTime}, ${totalTime}\n"

    f << learnLog


    NeedToReLearn = false

}
```

```
def getData(line, number) {

    //get data and convert to learnable format
    recordNo = number - 1
    def fmt_in = DateTimeFormat.forPattern("yyyy-MM-dd HH:mm:ss")
    def fmt_out = ISODateTimeFormat.dateTime()


    def splited = line.split(',')
    date = splited[0]
    logTime =  fmt_in.parseDateTime(date.replaceAll("\"", ""))
    day = "D" + ((logTime.getDayOfWeek() % 7) + 1)
    time = "P" + (logTime.getHourOfDay() + 1)
    rain_mm = splited[5]
    if(rain_mm.isNumber()) {
        def rainRange = ['T': 0.1, 'S':10.1, 'M':35.1, 'H':90.1, 'VH':Double.MAX_VALUE]
        rain = rainRange.find{rain_mm.asType(Double) < it.value}.key
    } else {rain = '?'}
    school = splited[4]
    roadLink = splited[134]
    connectedRoadLink1 = splited[134]
    connectedRoadLink2 = splited[3]

    roadLinkActual = roadLink


    //relearn condition
    if (learningSched != 0){
        if (recordNo%learningSched == 0) {
            NeedToReLearn = true
        }
    }

}
```

```
def randomMissing(missProb){

    switch(missProb) {
        case 0:
            break

        case 1:
            rain = '?'
            school = '?'
            connectedRoadLink1 = '?'
            connectedRoadLink2 = '?'
            roadLink = '?'
            break

        default:
            if(Math.random() <= missProb) rain = '?'
            if(Math.random() <= missProb) school = '?'

            if(Math.random() <= missProb) connectedRoadLink1 = '?'
            if(Math.random() <= missProb) connectedRoadLink2 = '?'
            if(Math.random() <= missProb) roadLink = '?'

    }

}
```

```
def infer() {

    //create instance
    Instances r = structure.stringFreeStructure()
    int classIdx=r.numAttributes() - 1
    r.setClassIndex(classIdx)
    def n = r.numAttributes()
    Instance inst = new Instance(n)
    inst.setDataset(r)

    def val = ['?',day,time,'?',rain,school,
                      connectedRoadLink1,connectedRoadLink2,roadLink]

    for(int i=0;i< n ;i++) {
        if(val[i] != '?') {
           inst.setValue(r.attribute(i),val[i].toString())
        } else {
           inst.setMissing(r.attribute(i))
        }

    }

    //choose matched model to classify

    def missableCTX

    switch (mode) {
        case 0:
            missableCTX = []
            break
        case 1:
            missableCTX = [5:rain, 6:school, 7:connectedRoadLink1, 8:connectedRoadLink2]
            break
        case 2:
            missableCTX = []
            break
        default:
            missableCTX = [7:connectedRoadLink1, 8:connectedRoadLink2]
    }

    //infer

    def inferStart= System.nanoTime()

    def matchedModel = "${name}.${algorithm}-" +
                            missableCTX.findAll{it.value == '?'}*.key.join()
    Classifier cls = (Classifier) weka.core.SerializationHelper.read("./data" +
                                  "/${name}${algorithm}/models/${matchedModel}.model")

    double infer = cls.classifyInstance(inst)
    roadLinkInferred = r.classAttribute().value((int)infer)


    def inferTime = (System.nanoTime() - inferStart)/1000
    if(roadLink == '?') totalInferTime += inferTime
```

```
    //Cal Accuracy for internal monitor, every records

    def point

    if(roadLink != '?') {

        recordWithActualVal ++

        if(roadLinkInferred == roadLink) {

            point = 1
            correct ++

        } else {

            point = 0
        }

        accuracy = (correct / recordWithActualVal * 100)

        currentCorrect << point
        if(currentCorrect.size > 10000) currentCorrect.remove(0)
        currentAccuracy = (currentCorrect.size == 10000) ?
                                                (currentCorrect.sum()/10000 * 100) : ''

    } else {

        point = ''
    }




    //Log: No,Log Time, Model, Traffic, Inferred, Point, Correct,
     //RecordWithActual, Accuracy, CurrentAccuracy
    def log = "$recordNo,$date,$matchedModel,"+
                "$roadLink,$roadLinkInferred,"+
                 "$point,$correct,$recordWithActualVal,$accuracy,$currentAccuracy\n"

    inferLog += log

    if(recordNo % 1000 == 0) {
        print '. '
        def inferFile = new File("./data/${name}${algorithm}/${name}inferLog.csv")
        inferFile << inferLog
        inferLog = ''
    }


}
```

```
def reportTraffic() {

    boolean activePeriod
    if(day == 'D1' | day == 'D7' | time == 'P24' | time == 'P1' | time == 'P2' |
                                time == 'P3' | time == 'P4' | time == 'P5' | time == 'P6' ) {
        activePeriod = false
    } else {
        activePeriod = true
    }


    //create report
    def report
    if(roadLink != '?'){

        report = "$date:,$roadLink,,,,,,,,\n"

    } else {

        recordWithMissingVal++
        if(activePeriod){
            recordWithMissingValA++
        } else {
            recordWithMissingValN++
        }

        def point

        if(roadLinkInferred == roadLinkActual) {

            point = 1
            reportCorrect++
            if(activePeriod){
                reportCorrectA++
            } else {
                reportCorrectN++
            }

        } else {

            point = 0
        }

        reportAccuracy = (recordWithMissingVal != 0) ?
                                    (reportCorrect / recordWithMissingVal * 100) : ''
        reportAccuracyA = (recordWithMissingValA != 0) ?
                                    (reportCorrectA / recordWithMissingValA * 100) : ''
        reportAccuracyN = (recordWithMissingValN != 0) ?
                                    (reportCorrectN / recordWithMissingValN * 100) : ''

        //Log: Log Time, Traffic, Inferred, Point, Actual,
          //Correct, Missing, Accuracy, Accuracy Active, Accuracy Non Active
        report = "$date:,,$roadLinkInferred,$point,$roadLinkActual," +
                "$reportCorrect,$recordWithMissingVal,$reportAccuracy," +
                "$reportAccuracyA,$reportAccuracyN\n"
    }

    //save
    reportLog += report
    if(recordNo % 1000 == 0) {
        def reportFile = new File("./data/${name}${algorithm}/${name}reportLog.csv")
        reportFile << reportLog
        reportLog = ''
    }


}
```

```
    def saveHistory() {

        //add data to arff

        def line = "\n$date,$day,$time,$rain_mm,$rain,$school,"+
                        "$connectedRoadLink1,$connectedRoadLink2,$roadLink"


        if(roadLink != '?') historyLog += line

        if(recordNo % 1000 == 0 | NeedToReLearn) {
            File f = new File("./data/${name}${algorithm}/${name}history.arff")
            f << historyLog
            historyLog = ''
        }


    }
```

```
    def summarize() {

        def inferFile = new File("./data/${name}${algorithm}/${name}inferLog.csv")
        inferFile << inferLog

        def reportFile = new File("./data/${name}${algorithm}/${name}reportLog.csv")
        reportFile << reportLog


        def f = new File("./data/${name}${algorithm}/${name}summarize.txt")

        def report =    "Road Link:                                $name\n"+
                        "Total record:                             $recordNo\n"+
                        "Sensor failure rate:                      ${missRate * 100}%\n"+
                        "Total missing actual traffic:             $recordWithMissingVal\n"+
                        "Total inferring time:                     $totalInferTime\n"+
                        "Total correct inferred traffic report:    $reportCorrect\n"+
                        "Total accuracy of inferred traffic report : $reportAccuracy\n"+
                        "\n----Internal Monitor----\n"+
                        "Overall correct inferred:                 $correct\n"+
                        "Overall accuracy:                         $accuracy\n"


        f << report


        def deployLog = new File("./final/deployLog.csv")
        deployLog << "${totalInferTime/recordWithMissingVal},$reportAccuracy," +
                         "$reportAccuracyA,$reportAccuracyN,"



    }

}
```

# Appendix C : Questions in Survey

## Questionnaire 20 Questions

The purpose of this online survey is for academic. In addition this research can be parts of improving the traffic report service. The survey is anonymous. No personal information will be disclosed or published. The approximate time to take is around 10-15 minute to complete the whole survey. However, you can quit at any stage of time if you feel uncomfortable or don't want to proceed. Thank you very much for helping us!

(Note: if the answer to this question is No, the survey will be terminated)

**Part1: Screening Question**

1 Do you drive a car in Bangkok as a routine manner?
1.Yes
2. No

**Part2:**

2 What is your gender?
1. Male
2. Female

3 What is your age range?
1. Lower than 18 years old
2. 18-25 years old
3. 26-33 years old
4. 34-41 years old
5. 42-49 years old
6. 50-57 years old
7. 58-65 years old
8. More than 65 years old

4 What is your highest level of education?
1. None
2. Primary School
3. Secondary School
4. Vocational School
5. Bachelor's Degree
6. Master's Degree
7. Doctorate's Degree

5 What is your marital status?
1. Unmarried
2. Married but no offspring
3. Married and have offspring
4. Widow/Divorce/Separated

6 What is your occupation?
1. Government Officer
2. State Enterprise Officer
3. Private Company Officer
4. Business Owner
5. Student
6. Others (Please specify)

7 Do you have your own car?
1. Yes
2. No

8 On average, how often do you drive a car per week?
1. 1-2 days/week
2. 3-4 days/week
3. 5-6 days/week
4. Everyday

9 Which area of Bangkok do you primarily stay?
1. Central Business District
2. Inner City
3. Outer City
4. Suburb of Bangkok
5. Nearby Provinces (Outside Bangkok)

10 To what extent do you drive for these following reasons?

| | Never | A bit | Medium | A lot | Main reason |
|---|---|---|---|---|---|
| A. Work/Study | 1 | 2 | 3 | 4 | 5 |
| B. Delivery offspring | 1 | 2 | 3 | 4 | 5 |
| C. Recreation | 1 | 2 | 3 | 4 | 5 |
| D. Travelling other provinces | 1 | 2 | 3 | 4 | 5 |
| E. Emergency | 1 | 2 | 3 | 4 | 5 |
| F. General Convenience | 1 | 2 | 3 | 4 | 5 |
| G. Reflect social status | 1 | 2 | 3 | 4 | 5 |

**Part 3: Bangkok Traffic**

11 Could you please give a short definition of "Bangkok Traffic"?

12 To what extent do you think the following factors affect the traffic in Bangkok?

| | No effect | A bit | Medium | A lot | Extremely effect |
|---|---|---|---|---|---|
| A. Specific time of the day (e.g. 6-7am, 11-12am,4-5pm,9-10pm) | 1 | 2 | 3 | 4 | 5 |
| B. Specific day of the week (Mon, Tue, Wed, Thu, Fri, Sat, Sun) | 1 | 2 | 3 | 4 | 5 |
| C. Group of the day (e.g. Weekday/Weekend) | 1 | 2 | 3 | 4 | 5 |
| D. Public holiday | 1 | 2 | 3 | 4 | 5 |
| E. Density of vehicles on the roads adjacent to the observed road | 1 | 2 | 3 | 4 | 5 |
| F. Level of falling rain | 1 | 2 | 3 | 4 | 5 |
| G. Some Incidents (e.g. Accident, Road construction, Special events, Protest | 1 | 2 | 3 | 4 | 5 |
| F. Others Please specify | | | | | |

**Part 4: Information Search for Knowing Traffic Condition**

13 Do you think how useful of knowing traffic condition toward your driving?
1. Not at all
2. A little useful
3. Not sure
4. Useful
5. Very useful

14 Which sources of traffic information do you usually use? (Multiple answers)

1. Traffic news from Television
2. Traffic news from Newspaper
3. Traffic report from General Radio
4. Traffic report from Traffic Radio
5. Electronic Board on the Road
6. Website showing traffic condition as a map
7. Website showing traffic condition as a video from video camera
8. Website showing traffic condition as a photo shot at a specific time
9. Website showing traffic condition as a map and a video from video camera
10. Using browser on mobile devices (e.g. mobile phone, tablet) that can access internet to access website showing traffic condition.
11. Application on mobile devices (e.g. smart phone, tablet) showing traffic condition
12. Ask around from Friends/Other people
13. Never search any information from any source

15 How often do you search for the following information?

| | Never | Rarely | Sometimes | Often | Every time |
|---|---|---|---|---|---|
| A1. Road/Route to the destination before starting the trip | 1 | 2 | 3 | 4 | 5 |
| A2. Road/Route to the destination during a trip | 1 | 2 | 3 | 4 | 5 |
| B1. Small road/A short-cut toward the destination before starting the trip | 1 | 2 | 3 | 4 | 5 |
| B2. Small road/A short-cut toward the destination during a trip | 1 | 2 | 3 | 4 | 5 |
| C1.Traffic Condition of the destination before starting the trip | 1 | 2 | 3 | 4 | 5 |
| C2.Traffic Condition of the destination during a trip | 1 | 2 | 3 | 4 | 5 |
| D1. Surrounded traffic condition of the destination before starting the trip | 1 | 2 | 3 | 4 | 5 |
| D2. Surrounded traffic condition of the destination during a trip | 1 | 2 | 3 | 4 | 5 |
| E1. Reasons/Causes of the traffic jam before starting the trip | 1 | 2 | 3 | 4 | 5 |
| E2. Reasons/Causes of the traffic jam during a trip | 1 | 2 | 3 | 4 | 5 |

16 If you know in advance that the traffic of one road in your route is going to be jammed definitely, what are you going to do?

1. Nothing. Be patient of such traffic jam because, no matter what happen, I have to pass that road.
2. Not doing anything yet because when I arrive there, the traffic should be better by then.
3. Decide to choose an alternative route. Searching information about other way I can go to the destination instead of the jammed one.
4. Cancel the trip on that day.

17 Imagine you are able to access and know the traffic condition from the following sources without any restriction. To what extent do you think the information from those sources will be useful for your driving?

| | Not at all | A bit | Not sure | Useful | Very useful |
|---|---|---|---|---|---|
| A. **Electronic Board,** showing traffic condition of **main roads**, as a **map with the colours** (e.g. green=no jam, yellow=a bit jam, red=very jam) | 1 | 2 | 3 | 4 | 5 |
| B. **Electronic Board,** showing traffic condition of **main roads** as a **real-time video** | 1 | 2 | 3 | 4 | 5 |
| C. **Radio,** reporting traffic condition of **main roads** from **relevant organization** | | | | | |
| D. **Radio,** reporting traffic condition of **main roads,** including **small roads and short-cuts** from **other drivers**. | 1 | 2 | 3 | 4 | 5 |
| E. **Application on mobile devices** (e.g. Mobile Phone, Tablet), showing traffic condition of **main roads** as a **map with the colours** (green=no jam, yellow=a bit jam, red=very jam) | 1 | 2 | 3 | 4 | 5 |
| F. **Application on mobile devices** (e.g. Mobile Phone, Tablet), showing traffic condition of **main roads,** including **small roads and short-cuts** as a **map with the colours** (green=no jam, yellow=a bit jam, red=very jam) and a real-time video | 1 | 2 | 3 | 4 | 5 |
| G. **Website,** showing traffic condition of **main roads** as a **map with the colours** (green=no jam, yellow=a bit jam, red=very jam) | 1 | 2 | 3 | 4 | 5 |
| H. **Website,** showing traffic condition of **main roads,** including **small roads and short-cuts** as a **map with the colours** (green=no jam, yellow=a bit jam, red=very jam) and a **real-time video** | 1 | 2 | 3 | 4 | 5 |

18 Normally, how often do you use/play social networks? (eg. Facebook, Twitter)

1. Never
2. Rarely
3. 1-3 days/week
4. 4-6 days/week
5. Everyday

**Part 5: Social Networks and Traffic Condition**

19 How often do you mention about the traffic condition in social networks?

1. Never
2. Rarely
3. Not sure
4. Many times
5. Often

**Part 6: Specific information of traffic condition mentioned in Social Networks**

20 If you mention about traffic on social network, how do you mention?

1. Only the traffic condition, but not including name of roads/places and time
2. Traffic condition and name of roads/places
3. Traffic condition and time
4. Traffic condition including name of the roads/places and time
5. Other (please specify)