Proceedings of the Eastern Asia Society for Transportation Studies, Vol. 5, pp. 1522 - 1532, 2005

TECHNIQUES FOR ANALYSING LARGE VOLUMES OF DATA GENERATED BY TRAFFIC RESEARCH

Nikolaos VOGIATZIS
PhD Candidate
Transport Systems Centre
University of South Australia, AUSTRALIA
GPO Box 2471
ADELAIDE, South Australia AUSTRALIA
Fax: +61 8 8302 1880
E-mail: nikolaos.vogiatzis@unisa.edu.au

Rocco ZITO
Deputy Director
Transport Systems Centre
University of South Australia, AUSTRALIA
E-mail: rocco.zito@unisa.edu.au

Hideto IKEDA
Professor, Cyberspace Technology
Department of Computer Science
Ritsumeikan University, JAPAN
E-mail: hikeda@cyber.cs.ritsumei.ac.jp

Frank PRIMERANO
Research Fellow
Transport Systems Centre
University of South Australia, AUSTRALIA
E-mail: frank.primerano@unisa.edu.au

Benjamin VANDERLINDE
Information Technology Support Officer
School of Natural and Built Environments
University of South Australia, AUSTRALIA
E-mail: benjamin.vanderlinde@unisa.edu.au

Waskitho WIBISONO
Masters Candidate
Department of Computer Science
Ritsumeikan University, JAPAN
E-mail: waskitho_w@yahoo.co.uk

Abstract: It is generally accepted that the volume of data being produced by transportation related applications is large. GIS, GPS and Urban Traffic Control Systems (UTC) all contribute a cornucopia of data which is ripe for analysis.

The Transport Systems Centre, like many other transport engineering centres, has in the past concentrated on using common desktop analysis tools to perform the necessary calculations. This approach requires the analyst to aggregate data to a point where they can manipulate it. However, with the volumes of data that are now routinely collected and analysed, it is becoming increasingly difficult to use these tools; they are not designed for such undertakings.

This paper discusses the needs of researchers dealing with transport data, and the steps towards designing a comprehensive research database.

Key Words: Database, Transportation Data, Spatial Database

1. INTRODUCTION

Increasingly, the use of computers in the analysis of transportation data is becoming important. There have been many articles in the past on this very point (Clement, 1995, Dia, 2001, Kecskemethy and Hiller, 1995, Vogiatzis, 2005), and the attempts by various research groups to design database schema's to achieve this. In this respect, the Transport Systems Centre is similar to all these other research groups. Researchers are routinely using simple database and spreadsheet packages to perform the analysis they require. However, such tools are not necessarily designed to handle the vast amounts of data generated within transportation research. As such, researchers are simply spending a great deal of time on the aggregation of data rather than its analysis.

The Transport Systems Centre at the University of South Australia along with the Cyberspace Technology Laboratory at the Ritsumeikan University is working towards developing a transport database that can be accessible from the internet and that acts as a data warehouse for transport.

The aim of this paper is to discuss the type of work being conducted in a transportation research organisation and the shift towards using more sophisticated tools to improve the speed, efficiency, and quality of the analysis being conducted. To do this, we will discuss the use of databases within traffic research, the need for the development of a purpose built transport research database, the core elements and design of such a database, the work that has been completed to date, followed by future work and conclusions.

2. DATABASES IN TRANSPORTATION RESEARCH

The use of databases in research is now common; they provide a mechanism by which large amounts of data can be analysed and visualised.

There are many types of databases and the term ‘database’ is so common that its use can easily be a source of confusion. Colloquially, a database is a collection of data that is made available to users. There is a search capability associated with the database and this allows a user to query the collection of data and return some sort of result. To a software engineer, however, a database is more than that; a database is also the ‘way’ in which the data is stored, the design/schema employed, and the way the data is made available. So a ‘database’ to a user is actually an ‘application’. In the technical sense of the term, a database can be as simple as a single sheet in a spread-sheet application; so we need to be somewhat specific about what is being discussed. For the purpose of this paper, we will freely interchange the terms ‘database’ and ‘database management system’ to be the same thing, and the ‘end-product’ used by users of the system will be referred to as the ‘application’; this will ensure we are consistent with the ‘information systems/software engineering’ terminology.

Database technologies have a thirty year history and that the types of technologies available differ towards the type of work that one wishes to perform, he also states that the most common type of database currently used within industry is the Relational Database Management System (RDBMS)(Ikeda, 2004). These systems use ‘relations’ to bind different data into groups and therefore provide a mechanism by which transactional information can be extracted(Slater, 1997, Vogiatzis, Ikeda, Woolley and He, 2003). These types of databases are designed to answer queries such as ‘how many widgets were sold in the previous quarter?’ or ‘what was the total amount of leave liability to the company year to date?’ Thus relational databases are not designed to model object behaviour, rather they are designed to model transactional relationships.

There has been a great deal of work to build related databases that can assist in transport planning. Systems such as Trafikplan (Taylor, 1992), and the work of Thong and Wong (1997) attest to this as does Choi and Jang (2000) and Clement (1995). The problem with such systems is that they use the relational data model, which is not always suitable for modelling transportation networks.

Ikeda and Vogiatzis (Ikeda and Vogiatzis, 2003, Ikeda, Vogiatzis, Wibisono and He, 2004, Ikeda, Vogiatzis, Wibisono, Mojarrabi and Woolley, 2004, Vogiatzis and Ikeda, 2003,

Vogiatzis *et al.*, 2003, Wibisono, Ikeda and Vogiatzis, 2004) along with a number of their colleagues are developing new types of database management systems (much of which we will discuss later) designed for managing large-scale/metropolitan-wide traffic management systems, whose primary purpose is to manage traffic in real-time using live/streaming data and historical data in the form of knowledge bases. This work is some years off completion and does not provide the mechanism by which one can analyse these vast data masses in an appropriate format now.

So, what other types of database systems are available? Some of the technologies available include (Ikeda, 2004):

- RDBMS (Relational Database Management Systems) as already discussed;
- OODBMS (Object Oriented Database Management Systems) which we will discuss in a moment;
- Distributed Database Technologies: database management systems that operate over many computing resources;
- Data warehousing: used to provide some form of 'business intelligence' from multiple sources, primarily in the form of statistical intelligence; and
- Streaming database technologies: used for collecting and analysing data as it enters the database, typically used in computer network monitoring systems.

3. NEXUS: TRANSPORT RESEARCH SYSTEM

NEXUS is an application being developed by Vogiatzis (2004) in collaboration with the authors of this paper as a reporting system for transport research. The objectives of this system are simple: devise and develop an internet accessible database application that allows researchers to data mine transportation data.

This statement is somewhat simplified; we will now begin to discuss the reality.

Traditionally, the nature of traffic modelling focused on issues related with the 'Level of Service' (LOS) offered by a road section or intersection. The modelling would test various scenarios which would provide some level of advice as to whether changes proposed would affect the flow of traffic. This type of modelling rarely, if ever, specifically took into account issues such as emissions or the economic impact of changes; this does not suggest that such research was not conducted. It was difficult to do so due to the types of tools typically used within transportation research and the limited access to data that was being produced; resulting in a lack of rigour with many models. We have now access significantly more data and better tools for analysis, for example SCATS produces historical VS (traffic volume data as detected by each detector at an intersection) and SM (strategic monitor records; these provide information relating to the signal phase that has been set, and the status of the traffic flow, such as 'oversaturated flow' etc) files for either five or fifteen minute blocks (McCabe, 2004); thus there is a need for better data/information management techniques to be applied.

However the improvements in computing, and the ability to collect more and more data has now provided the mechanism by which researchers can begin harnessing the data that is collected. For this to be of ultimate value to researchers there is a need to build an object model of transportation. From there one needs to devise a database schema that will allow for

data mining.

The birth of NEXUS was simple; a relational database that could analyse millions of records generated by UTC (Urban Traffic Control) systems to provide researchers with phase and volume information quickly. The prototype is already returning benefits with phase and volume data being provided for a CBD traffic microsimulation model of Adelaide, South Australia and for emissions modelling. A sample of the output can be seen in figure 1 and table 1.

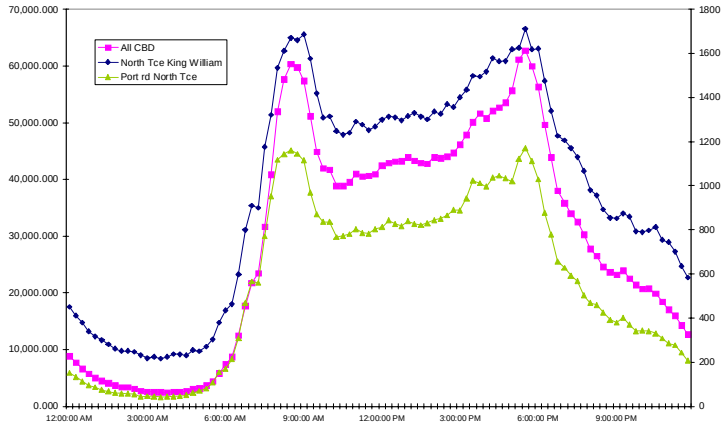

Figure 1 Traffic Volume Report from NEXUS using Adelaide CBD SCATS data; five million raw data points aggregated down to 96 data points within ten minutes, analysed within one hour

Table 1 Correlation of data using NEXUS report

	North Tce - King William St	Port Rd - North Tce	All CBD
North Tce - King William St	1.00		
Port Rd - North Tce	0.98	1.00	
All CBD	0.99	0.98	1.00

However severe limitations have been identified with the NEXUS project direction and this was: 1) using the relational model did not allow researchers to bind the volume and signal phase data with other transport objects; 2) the analysis only provided limited information (it could only give volumes and phases for intersections and the whole CBD collectively, but additional analysis from other data sources was required before these reports could provide true insight; and 3) the RDBMS chosen did not allow for simple management of the data itself. Unrelated to the RDBMS chosen was the fact that the server filesystem was inappropriate for the volume of data being collected and analysed and thus this caused the RBDMS to fail once 90 million records were inserted into a specific table.

The database used for the prototype is MySQL v4.0.18 production release, on a RedHat 8.0 server (really a 'big' PC) using the ext2 filesystem. Although for general purposes, this configuration would have been sufficient, it was quickly discovered that it was deficient in a number of key areas:

- The use of the **ext2** filesystem meant that files could not be created that were greater than 2 GB in size (90 million 'VS' SCATS records in a singular table is approximately 2 GB in size), this limitation is only on files that are created as a part of some application, it does not mean that files larger than 2 GB can not be stored on the server;

- MySQL v4.0.18 does not have stored procedure/trigger functionality (although v5.0 does have a stored procedure facility, for the uploading of a large number of records it is more appropriate to have a trigger facility as would be available in v5.1, which is several years away from becoming a production tool);
- The relational only nature of the RDBMS does not allow for 'objects' to be created.

In order to avoid some of the deficiencies, the database was segregated so as to ensure that table sizes of less than 2 GB were created and the use of Java to compensate for the lack of stored procedure and trigger facilities was employed.

A number of significant issues with this process were identified and to remedy the immediate problems a two stage operation was employed. The parsed data from the raw VS files were uploaded as CSV (comma separated values) file into a temporary table. Interestingly, the uploading process would on average take one minute per million records. This temporary table was then used by a Java programme to convert the parsed data into a usable format within the database. This process was significantly longer, with several hours required to complete the VS records and several days for the SM records.

Once the data was loaded, as detailed, it is a simple process to produce reports based on these millions of records, for all intersections and for all detectors, within an extremely short space of time (ten minutes to perform all the aggregations followed by one hour for the report in figure 1).

This however is not sufficient; in order to achieve true gains, one needs to be able to combine the data from numerous data sources, including land-use policy, emissions studies, crime data, and naturally, traffic volumes.

This suggests that the type of database that is of interest is a hybrid object-oriented and relational database. In some instances we will be interested in the 'transactional' nature of the data and in other instances we will be interested in the 'object' nature of the data.

4. WHAT ARE THE CORE ELEMENTS

There have been many attempts at designing an object model for transportation networks; however most of them tend to overly simplify the models used. Vogiatzis (2005) begins the process of developing a holistic object and component model for transport systems. He discusses the need to identify more than just network objects at a superficial level; that it is important to model a transport network from the highest level first. This ensures that the one can also include the transactions and relationships that occur between objects; as much as the data associated with each entity. Figure 2 shows a draft and incomplete component model for transport networks. This shows the beginnings of a comprehensive model that will be used for modelling transport objects within the context of Locality-Scope and its associated information systems.

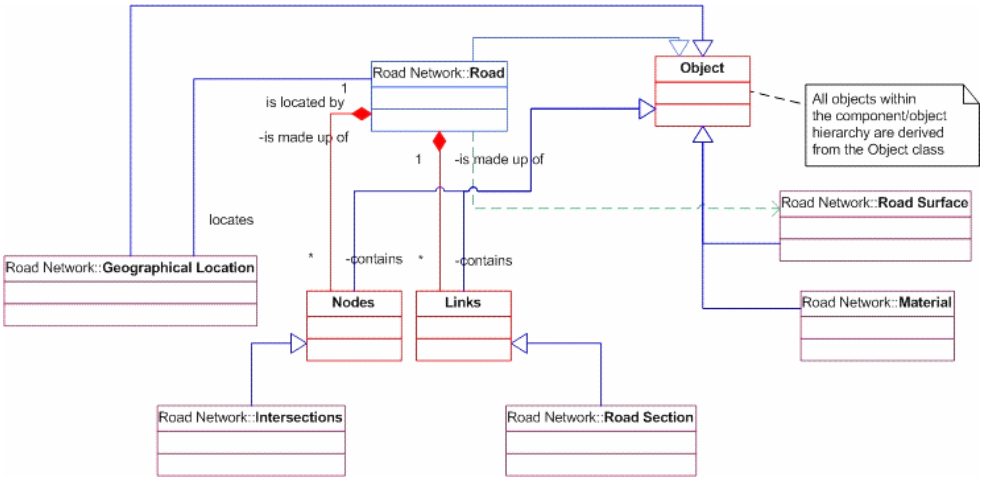

Figure 2 Draft and incomplete component model of transportation networks (Vogiatzis, 2005)

The Transport Systems Centre has purchased computing cluster for the task of analysing large volume sets and is in the process of designing the necessary database schema for this task.

There are two main components of the system; the first is the hardware cluster components which include the operating system and the backend database as below:

Table 2 Items list for the Transport Systems Centre Cluster (Dynamis)

Item
Intel P4 3.2 GHz Motherboards
2 Gb RAM
1 x 80 Gb SATA HDD (Operating System and software)
1 x 250 Gb SATA HDD (data)
Microsoft® Windows™ 2003 Enterprise Edition
ORACLE 10g Server
Rack mountable

There are five nodes (computers) within the cluster; each of which works independently and whose resources (CPU/HDD) can be bound together for the purpose of high computational needs.

The software components on the other hand are a combination of off the shelf software solutions and custom software developed by members of the Transport Systems Centre at the University of South Australia or of the Cyberspace Technology Laboratory at the Ritsumeikan University, Japan.

For custom software, all components and libraries will be written using the .NET platform and access to the resource will be through a cluster controller which is being designed to provide profiling and accounting. The system is being designed to allow users to access services and databases through a suite of custom libraries ensuring that no user has direct access to the cluster and to ensure that cluster 'jobs' are prioritised.

In addition, the custom software component is itself from three separate components; one component is the backend database with TSC developed 'stored procedures', the middle 'layer' also known as the 'business logic' layer, whilst the other is 'client' software used to

access the database externally.

4.1. System Architecture

This reporting system is useful for more than just managing traffic however; it is also useful for more strategic traffic/transport network analysis as is detailed in section 5.

To achieve this, the database of choice is ORACLE 10g; this database management system allows for the design and development of Object-Relational databases. An object-relational database is one that allows for the design of object-oriented databases and relational databases; effectively allowing the system designer access to the best of both 'worlds'. There are many other such database management systems in existence; however ORACLE has a long history of operation and the software developers have most experience with this particular product. Other database management systems that could have also been used include: DB2, Microsoft SQL Server, Informix, and SAP. ORACLE also has a Spatial Data engine; this though is common with many database management systems including MySQL and the enterprise database engines above having some form of spatial engine

According to Vogiatzis (2005) the component model is the key to developing such a system; however he does not cover the relational aspect of the system. This is not necessarily an error of judgement, rather an intentional desire to design and built an object/component model that models transportation as accurately as possible where within an object/component model the relational model is a component. However from the perspective of a database application, one needs to be explicit about the transactional nature of events.

Much of the data available is designed for transactional systems; as such it is not always possible to include it directly into an object database. As such, it is practical to import and manage data in a relational manner, and then manipulate it into an object model (where appropriate).

5. APPLICATIONS AND WORK COMPLETED

There are numerous applications that one can use from even a 'cut-down' version of NEXUS. In the first instance, NEXUS 1.0 will simply be an extension of the NEXUS 0.1 prototype with the inclusion of emissions and GIS data.

As one can see from figure 3, the design encompasses all aspects of transportation research; from the design of a new generation intelligent (not adaptive) traffic network management system such as IMAGINATION (Vogiatzis *et al.*, 2003), to emissions modelling projects such as NISE2 (National In-Service Emissions Study 2: Australia), and for City Logistics(Tseng, 2004); this database forms the basis of many applications.

It is an important aspect of the design of the database; to be the core repository for data from many different sources, and thus allow for reporting, and ultimately network management, to be performed with as a complete a picture as is possible.

The work relating to the development is being conducted on two 'fronts': one is the specific work relating to the design and development of the database and the other is the research

projects that are driving the design requirements.

From the perspective of the database design and development a number of tasks have been completed thus far:

- The prototype NEXUS reporting system has been completed and proven functional (see figure 1 above);
- The emissions modelling database conversion planning has commenced;
- Planning and design has commenced on the full version NEXUS that will reside on the Dynamis platform which combines the SCATS traffic management system data and emissions modelling;
- The cluster has been installed and configured ready for use;
- The conceptual fundamentals of the Locality-Scope model have been completed;

As mentioned, there are a number of research projects driving the development of this database and they include:

Table 3 Projects that use/will use the Dyanmis/NEXUS system

Project Name	*Category*	*Description*
PhD Programme (Nikolaos Vogiatzis)	Post-graduate	Titled *"On the Locality-Scope Model, Data Mining and the Application of Advanced Techniques for Improved Transportation Network Management Integration"* this work focuses on the development of a new generation intelligent traffic management system. It does so by demonstrating the need to data mine existing traffic/transport databases and through the development of 'Locality-Scope', a new way of modelling systems.
Australian Research Council Project (ITS Implementations)	Australian Government Funded	This research project investigates the advantages and disadvantages of using Intelligent Transport Systems (ITS) technologies in an Australian Central Business District (CBD) environment. It is often assumed that road networks and especially high activity areas such as CBDs can benefit significantly from ITS implementations. Traditionally ITS impacts have been difficult to quantify due to difficulties in isolating network effects and limitations with traditional traffic models. The project overcomes this difficulty by using a current state of the art traffic microsimulation model in order to test various ITS scenarios in a duplicate of a real world CBD (Woolley, Taylor and Hewitt, 2002).
3LOM/Locality-Scope Database Management System	Collaborative	This work focuses on the development of a new database management system using Locality-Scope as the main premise of operation. This research is being conducted in parallel with the implementation of 3LOM using more conventional database management systems.
Bistatic Radar in Transport	Collaborative	This project looks at the use of Bistatic Radar to perform Origin-Destination and Route Choice Surveys automatically, 24 hours a day, 7 days a week (Vogiatzis, Mojarrabi, Ikeda, Kubik and Mojarrabi, 2005)
National In Service Emissions Study 2 (NISE2)	Australian Government	This project focuses on the development of emissions profiles for petrol fuelled vehicle under

Project Name	*Category*	*Description*
	Funded	3.5 tonnes to assist in the modelling of green-house gases (Primerano and Zito, 2004).
The advancement of a road driving assessment tool	UniSA funded	This project looks at redeveloping a driving assessment tool to assist in deriving increased tool performance, and more sophisticated report generation (Zito and Primerano, 2004).
PhD Programme (Yung-yu Tseng)	Post-Graduate	This work focuses on the reduction of unnecessary freight movement within an urban environment. This will be done using the principles of City Logistics (Tseng, 2004).

From the project list in table 3 it is easy to note that all these projects require the same type of data: traffic signal volume and phase settings, vehicle performance (emissions, etc), and route choice and origin-destination data. Therefore, it is appropriate that not only a singular repository should be developed, but that repository should provide the mechanism by which data can be viewed in 'different' ways by the different groups.

NEXUS on the Dynamis platform is being designed with this in mind, the ability to share data amongst various users, and making available a rich source of data for in-depth analysis.

6. FUTURE WORK/RESEARCH

In the immediate future, the team will focus their attention on the migration of their existing databases into an intermediary database application to enhance the immediate processing capabilities of their respective systems.

Then the team will begin the design of a singular system. This will require the team to identify common ground between all the projects mentioned above and to design the appropriate database schema and application system that will allow for reporting based on the various resources available to them.

This opens the door to new and innovative ways to analyse and visualise the data that is available.

7. CONCLUSIONS

Traditionally teams of researchers have developed their own unique systems for their research at hand. The Transport Systems Centre at the University of South Australia and the Cyberspace Technology Laboratory of the Department of Computer Science at Ritsumeikan University has begun a long-term collaboration aimed at:

- Designing a Component/Object model of the entire transport system;
  - This includes road networks and transport activities;
- Designing and implementing an Object-Relational database that will form the basis of the middle and top layer of the 3LOM (Three Layer Object Model for the Integration of Transportation Systems);
- Implementing an immediate intermediary solution for the improved reporting of transport related analysis;
- The application of this database system to the modelling of ITS implementations, emissions modelling and city logistics modelling.

It is hoped that this collaboration will progress the discipline of Transport Information Science.

REFERENCES

Choi, K. and Jang, W. (2000) Development of a transit network from a street map database with spatial analysis and dynamic segmentation,**Transportation Research Part C: Emerging Technologies**, **Vol. 8**, **No. 1-6**, 129-146.

Clement, S. (1995) *The Transport Network Relational Database*, University of South Australia, Adelaide, Australia, Working Paper 95/1

Dia, H. (2001) An object-oriented neural network approach to short-term traffic forecasting,**European Journal of Operational Research**, **Vol. 131**, **No. 2**, 253-261.

Ikeda, H. (2004) Personal Communication - On database technologies, Vogiatzis, N.,

Ikeda, H. and Vogiatzis, N. (2003) Personal Communications - On Solving the Information Latency Problem in IMAGINATION, Vogiatzis, N.,

Ikeda, H., Vogiatzis, N., Wibisono, W. and He, Y. (2004) An Algorithm for Reducing Detection Load in Transportation Oriented DSMS, **Proceedings 27th Australasian Transportation Research Forum**,Adelaide, South Australia,

Ikeda, H., Vogiatzis, N., Wibisono, W., Mojarrabi, B. and Woolley, J. E. (2004) Three Layer Object Model for Integrated Transportation System, **Proceedings 1st International Workshop on Object Systems and Systems Architecture**,Victor Harbor, Australia, 11-14 January 2004

Kecskemethy, A. and Hiller, M. (1995) Object-oriented programming techniques in vehicle dynamics simulation,**Mathematics and Computers in Simulation**, **Vol. 39**, **No. 5-6**, 549-558.

McCabe, G. (2004) SCATSSIM: Interfacing SCATS and Traffic MicroSIMulations, **Proceedings Workshop on Traffic Simulation**,Brisbane, Australia,

Primerano, F. and Zito, R. (2004) NISE2 -Contract 2 Drive Cycle Development Methodology, 'Dec 2004'

Slater, D. (1997) Your Database is About to Become More Complicated, 'Nov 1

Taylor, M. A. P. (1992) **Trafikplan**, School of Engineering, University of South Australia, Adelaide, South Australia

Thong, C. M. and Wong, W. G. (1997) Using GIS to design a traffic information database for urban transport planning,**Computers, Environment and Urban Systems**, **Vol. 21**, **No. 6**, 425-443.

Tseng, Y.-y. (2004) Exploring City Freight Logistics Problems, **Proceedings 26th Annual Conference of Australian Institutes of Transport Research**,Melbourne, Australia,

Vogiatzis, N. (2004) **NEXUS: Transport Research System**, Transport Systems Centre, Adelaide, Australia

Vogiatzis, N. (2005) Objects and Components in the Locality-Scope Model (Full Draft Submitted), **Proceedings Intelligent Vehicles and Road Infrastructure (IVRI05)**,Melbourne, Australia, 16th & 17th of February 2005

Vogiatzis, N. and Ikeda, H. (2003) Personal Communications - On Solving the Information Latency Problem in IMAGINATION, Vogiatzis, N.,

Vogiatzis, N., Ikeda, H., Woolley, J. and He, Y. (2003) Integrated Multi-Nodal Traffic Network Systems,**The Journal of the Eastern Asia Society for Transportation Studies**, **Vol. 5**, **No.**, 2092-2107.

Vogiatzis, N., Mojarrabi, B., Ikeda, H., Kubik, K. and Mojarrabi, B. (2005) Problems Associated with OD and Trip Choice Surveys and the Shift Towards Galileo for the Solution (Abstract Submitted), **Proceedings Eastern Asia Society for Transportation Studies**,Bankok, Thailand,

Wibisono, W., Ikeda, H. and Vogiatzis, N. (2004) Multi-purpose Front-End devices for Integrated Transportation Systems, **Proceedings Seminar Nasional Ilmu Komputer dan Teknologi Informasi (SNIKTI 2004)**,Indonesia,

Woolley, J. E., Taylor, M. A. P. and Hewitt, P. (2002) *Australian Research Council Linkage - Projects (Round One) Application Form for Funding Commencing 2003: Traffic Microsimulation of ITS Implementations in CBD road networks*, University of South Australia and the Adelaide City Council, Adelaide, South Australia,

Zito, R. and Primerano, F. (2004) *The advancement of a road driving assessment tool*, University of South Australia, Adelaide, Australia, Small Grants Scheme 2005