Research fellowship – Final report
Candidate: Yang Bang-hua

Yang Bang-hua has worked on the following topics during her stay at IMOB

1) Literature study about activity-based modeling

Modelling traffic patterns / travel demand has always been of vital importance in the domain of transportation science. Due to the rapid increase in worldwide car ownership and use; several models of transport mode, route choice and destination have been used by transportation scientists. These models are necessary to predict travel demand on the long run and to support investment decisions in new road infrastructure. Initially, travel was assumed to be the result of four subsequent decisions that were modelled separately.

A well-known and widely used method to replicate and simulate behaviour, is the four-step modelling approach (Ruiter & Ben-Akiva, 1978). A four-step model is a trip-based model including a series of mathematical models calibrated on origin-destination data (McNally, 2008). On the one hand, a behavioural confrontation with regard to the use of these types of modelling exists, nevertheless it is still very popular and applied frequently due to its simplicity and limited computational requirements. On the other hand, four-step modelling contains a number of limitations as formulated in (McNally, 2000). As trip-based models focus on individual trips, they neglect the basic principle that the demand for travel is derived from activity participation, and that trips / activities are spatially and temporally related. Furthermore, four-step models neglect certain behavioural aspects such as complex choice sets, household dynamics and interrelationships between travel and activity participation.

As a next step in travel-demand modelling, a move was made towards tour-based modelling that is assumed to enhance the behavioural realism by combining the series of separate trips into tours. Such type of modelling is established on the idea that each tour contains one main goal i.e. the primary destination. The trip to and from this primary destination can be interrupted by a number of intermediate stops. To identify the primary destination in tour-based models has the following purposes; (i) fixed activities (work or school), (ii) maintenance activities (shop or pick-up/drop-off) and (iii) flexible activities (social, leisure, other).

Activity-based models recently gained more importance although tour-based models offer greater behavioural realism as compared to the four-step models. Instead of concentrating on the connection between the activities executed in the course of the same home- or work-based tour, the activity-based models focus on the relationship between all activities executed in the course of a day/week/month with the interactions between household members. Activity-based models involve the application of well-developed theoretical constructs to empirical data, with the aim of generating a predictive model that can be used for generalization purposes and for the evaluation of Travel Demand Measures (TDM). Activity-based models can be subdivided into; (i) constraints-based, (ii) simultaneous and (iii) computational process models.

Constraints-based models examine whether particular activity patterns can be realized within a specified time-space environment (Timmermans et al., 2002). The space-time environment is defined in terms of locations, their attributes, available transport modes and travel times between locations per transport mode. Examples of these types of models are the Lenntorp's (1976) PESASP model, the CARLA model (Jones et al., 1983), the MASTIC (Dijst, 1995; Dijst & Vidakovic, 1997) model and Kwan's (1997) GISICAS model. In comparison to other known models, constraints-based models lack the necessary mechanisms to predict adjustment behaviour of individuals. Furthermore in such models, policies may often have less dramatic social impacts.

Simultaneous models are often based on the assumption of utility-maximising behaviour. Individuals are assumed to schedule their activities such that their utility is maximized. The theory is based on the assumption that choice alternatives can be represented as bundles of attribute levels, for which a particular utility can be derived. Constraints are usually not included in much detail. Activity scheduling behaviour is not addressed specifically in these models but follows automatically from the prediction of the full activity-travel patterns (Timmermans, 2001). The nested logit formulation became the most frequently applied technique in simultaneous activity-based models of transport demand. Several examples can be given but the STARCHILD model (Recker et al., 1986a; Recker et al.,1986b), the STPG model (Kitamura et al., 2000), the daily activity schedule program (Ben-Akiva et al., 1996; Bowman, 1995; Ben-Akiva & Bowman, 1995; Bowman et al., 1998; Bowman & Ben-Akiva, 1999) and the the PETRA model (Fosgerau, 1998) are well-known examples. In addition to these nested logit models, there are other attempts for capturing decision and scheduling behaviour by means of utility-maximizing theories. Examples are the PCATS model (Kitamura & Fujii, 1998), the work by Recker (1995), by Bhat and Singh (2000) by and Bhat (1999) and by Bhat and Misra (2001, 2002). In addition to this, Bhat also suggested a series of models to predict more separated components of the activity scheduling decision (Bhat, 1996, Bhat & Singh, 1997). These different models came together in the CEMDAP model (Bhat et al., 2004), that is the only operational model in this category.

Computational process models have received increased attention in the last couple of years because utility-maximizing models do not always reflect the true behavioural mechanisms underlying travel decisions. The argument is that people may reason more in terms of context-dependent IF-THEN-ELSE structures when faced with different constraints and circumstances than in terms of truly maximizing utility-based behaviour. For this reason, several studies have shown an increasing interest in computational process models in order to model activity-diary data. Several examples of such models are the SCHEDULER computational process model, developed by Gärling et al. (1989), the AMOS model (Pendyala et al., 1995; Pendyala et al., 1998; Kitamura et al., 1995; Pendyala et al., 1997), the SMASH model (Ettema et al., 1994; Ettema et al., 2000), and the operational Albatross model *(A Learning-Based Transportation Oriented Simulation System)* model developed by Arentze and Timmermans (2000, 2005). Recently, another model has been made operational for the State of Florida under the name FAMOS: *Florida's Activity Mobility Simulator* (Pendyala, 2004).

Lastly, a research programme coordinated by Transportation Research Institute (IMOB) at Hasselt University, was initiated to develop a prototype, activity-based model of transport demand for Flanders (Bellemans et al., 2010; Janssens & Wets, 2005). The basis of this model, which has been given the acronym FEATHERS *(Forecasting of Evolutionary Activity-Travel of Households and their Environmental RepercussionS).* The environment is established as the spatio-temporal aggregate where an agent lives and executes its own daily schedule.

Literature Review:

Arentze T.A., and Timmermans H.J.P. 2000. ALBATROSS: A learning-based transportation oriented simulation system. Eindhoven University of Technology. EIRASS.

Arentze, T.A., and H.J.P. Timmermans. 2005. "The sensitivity of activity-based models of travel demand: results in the case of Albatross." In A. Jaszkiewicz, M. Kacmarek, J. Zak, M. Kubiak(ed.): Advanced OR and AI methods in Transportation. 573 – 578.

Bellemans T., B. Kochan, D. Janssens, G. Wets, T. Arentze, and H. J. P. Timmermans. 2010. Implementation Framework and Development Trajectory of the Feathers Activity-Based Simulation Platform. In Transportation Research Record: Journal of the Transportation Research Board, No. 2175, Transportation Research Board of the National Academies, Washington, D.C., pp. 111-119.

Bhat, C.R. 1996. A hazard-based duration model of shopping activity with nonparametric baseline specification and nonparametric control for unobserved heterogeneity. Transportation Research B, 30, 189-207.

Bhat, C.R. 1999. A comprehensive and operational analysis framework for generating the daily activity-travel pattern of workers. Paper presented at the 78th Annual Meeting of the Transportation Research Board, Washington, D.C., USA.

Bhat, C.R. and Misra, R. 2001. A comprehensive activity-travel pattern modelling system for non-workers with empirical focus on the organization of activity episodes. Paper presented at the 80th Annual Meeting of the Transportation Research Board, Washington, D.C., USA.

Bhat, C.R. and Misra, R. 2002. Comprehensive activity-travel pattern modeling system for non-workers with empirical focus on the organization of activity episodes. Transportation Research Record, 1777, 16-24.

Bhat, C.R. and Singh, S. 1997. A joint model of work mode choice, evening commute stops and post-home arrival stops. Final report, submitted to U.S. DOT Region 1, MIT.

Bhat, C.R. and Singh, S. 2000. A comprehensive daily activity-travel generation model system for workers. Transportation Research A, 34, 1-22.

Bhat, C.R., Guo, J.Y., Srinivasan, S. and Sivakumar, A. 2004. Comprehensive econometric microsimulator for daily activity-travel patterns. Transportation Research Record, 1894, 57-66.

Bowman, J.L. and Ben-Akiva, M.E. 1999. The day activity schedule approach to travel demand analysis. Paper presented at the 78th Annual Meeting of the Transportation Research Board, Washington, D.C., USA.

Bowman, J.L., Bradley, M., Shiftan, Y., Lawton, T.K. and Ben-Akiva, M.E. 1998. Demonstration of an activity-based model system for Portland. Paper presented at the 8th World Conference on Transport Research, Antwerp, Belgium.

Bowman, John L. 1995. Activity Based Travel Demand Model System with Daily Activity Schedules. Master thesis, Massachusetts Institute of Technology.

Dijst, M. 1995. Het Elliptisch Leven. Ph.D dissertation, KNAG, Utrecht University, Utrecht.

Dijst, M. and Vidakovic, V. 1997. Individual action space in the city. In Ettema, D.F., Timmermans, H.J.P. (Eds.): Activity-Based Approaches to Activity Analysis, Pergamon Press, Oxford, 73-88.

Ettema, D.F., Borgers, A.W.J. and Timmermans, H.J.P. 1994. Using interactive computer experiments for identifying activity scheduling heuristics. Paper presented at the 7th International Conference on Travel Behavior, Santiago, Chile.

Ettema, D.F., Borgers, A.W.J. and Timmermans, H.J.P. 2000. A simulation model of activity scheduling heuristics: an empirical test. Geographical and Environmental Modelling, 4, 175-187.

Fosgerau, M. 1998. PETRA: an activity based approach to travel demand analysis. Paper presented at the 8th World Conference on Transport Research, Antwerp, Belgium.

Janssens, D. and Wets, G. 2005. The presentation of an activity-based approach for surveying and modelling travel behaviour. Proceedings of the 32nd

Colloquium Vervoersplanologisch Speurwerk 2005: Duurzame mobiliteit: “Hot or not?”, editie 32,deel 7, Antwerp, Belgium, 1935-1954. Paper received award for most innovative paper.

Jones, Peter M., Dix, Martin C., Clarke, Mike I., & Heggie, Ian G. 1983. Understanding Travel Behaviour. 1st edn. Aldershot, England: Gower Publishing Company Limited.

Keren, D., Yasar, A., Knapen, L., Cho, S., Bellemans, T., Janssens, D., Wets, G., Schuster, A., and Sharfman, I. 2012. Exploiting graph-theoretic tools for matching and partitioning of agent population in an agent-based model for traffic and transportation applications, In the proceedings of ABMTRANS’12, Procedia Computer Science (Niagara Falls).

Kitamura, R. and Fujii, S. 1998. Two computational process models of activity-travel choice. In Gärling, T., Laitila, T. and Westin, K. (Eds.): Theoretical Foundations of Travel Choice Modelling, Elsevier, Oxford, 251-279.

Kitamura, R., Pendyala, R.M., Pas, E.I. and Reddy, D.V. 1995. Application of AMOS: an activity-based TCM evaluation tool applied to Washington D.C. metropolitan area. Proceedings of the 23rd Summer Annual Meeting, London, United Kingdom, 177-190.

Kwan, M.P. 1997. GISICAS: an activity-based travel decision support system using a GIS interfaced computational process model. In Ettema, D.F. and Timmermans, H.J.P. (Eds.): Activity Based Approaches to Activity Analysis, Pergamon Press, Oxford, 263-282.

McNally, Michael G. 2000 (Dec.). The Activity-Based Approach. Tech. rept. UCI-ITS-ASWP-00–4. Center for Activity Systems Analysis., Irvine, California.

McNally, Michael G. 2008 (Nov.). The Four Step Model. Tech. rept. UCI-ITS-AS-WP-07–2.Center for Activity Systems Analysis, Irvine.

Pendyala, R.M. 2004. FAMOS: application in Florida. Paper presented at the 83rd Annual Meeting of the Transportation Research Board, Washington, D.C., USA.

Pendyala, R.M., Kitamura, R. and Reddy, D.V. 1995. A rule-based activity-travel scheduling algorithm integrating neural networks of behavioural adaptation. Paper presented at the EIRASS Conference on Activity-Based Approaches, Eindhoven, The Netherlands.

Pendyala, R.M., Kitamura, R. and Reddy, D.V. 1998. Application of an activity-based travel demand model incorporating a rule-based algorithm. Environment and Planning B, 25, 753-772.

Pendyala, R.M., Kitamura, R., Chen, C. and Pas, E.I. 1997. An activity-based microsimulation analysis of transportation control measures. Transport Policy, 4, 183-192.

Recker, W.W., McNally, M.G. and Root, G.S. 1986a. A model of complex travel behavior: Part 1: theoretical development. Transportation Research A, 20, 307-318.

Recker, W.W., McNally, M.G. and Root, G.S. 1986b. A model of complex travel behavior: Part 2: an operational model. Transportation Research A, 20, 319-330.

Ruiter, E.R. and Ben-Akiva, M.E. 1978. Disaggregate travel demand models for the San Francisco bay area. Transportation Research Record, 673, 121-128.

Timmermans, H.J.P. 2001. Models of Activity Scheduling Behaviour. In: Stadt Regional Land, Vol. 71, 63-78.

Timmermans, H.J.P., Arentze, T.A. and Joh, C.-H. 2002. Analyzing space-time behavior: new approaches to old problems. Progress in Human Geography, 26, 175-190.

2) Statistical analysis of the BELDAM data and getting to know the FEATHERS framework. Making FEATHERS ready for Belgium.

The candidate was provided with the BELDAM data and performed statistical analyses on that data. Furthermore the candidate was introduced to the FEATHERS framework so that she could work independently in order to obtain an activity-based model for the BELDAM study area (Belgium). In order to run the FEATHERS activity-based scheduler for the Belgian situation, several data layers inside the FEATHERS database system were prepared. Activity-based schedule information, a synthetic population data set and environment information about the study area in terms of zoning system, land use and transportation system were processed. In the following sections the different steps involved in the data processing together with additional background information about the different types of data sets are provided.

Schedule data

Activity-based models differ highly from traditional transport forecasting models in the sense that the former models aim at predicting the interdependencies and interrelationships between the multitude of facets of activity profiles on an individual level. The major distinction with conventional models is that scheduling of activities comprises the foundation of activity-based models. Therefore, and in line with the basic underpinnings of the activity-based paradigm, the data required to estimate an activity-based model differs from the data required to build conventional models. More specifically, in order to build an activity-based model of transport demand, data on activity patterns are required. Given the needs of the activity-based modelling approach, the travel survey to be called in has to pay attention on the measurement of activities at the end of trips and to how and when the respondent chose to do them. One travel survey for the Belgian study area that can be used for estimating the activity-based model inside FEATHERS is the BELDAM travel survey. This BELDAM survey formally is a trip-based survey, however information about trip purposes and hence information about activities in between trips is available. Therefore, this survey is particularly suitable for estimating the activity-based model embedded in the FEATHERS framework. The BELDAM travel survey was conducted through 15.888 persons that were selected based on a random sample from the national register. These persons were all involved in a survey where information about the demographic, socioeconomic, household and trip-making characteristics of these individuals were gathered and for the purpose of this research, all person records and their according travel were then processed and being used as input for estimating the activity-based model incorporated inside FEATHERS.

Synthetic population data

Activity-based models require detailed information on household and person demographics and characteristics. Because of the fact that in Flanders, the gathering of individual data, or the retrieval of individual data from administrative registers is not allowed for privacy reasons, some missing attributes describing the population still had to be estimated. Therefore, a synthetic population data set had to be generated that represented households and household members. A synthetic population is meant to be a statistically duplicate of an actual population. For each household, characteristics such as number of household members, yearly income, number of cars, etc. were generated. Subsequently, each person got characterized by means of attributes such as age, gender, work status and driving licence. In this study, the aim was to create a synthetic population for Belgium. An application of Beckman et al. and Guo and Bhat's approaches for generating synthetic populations was chosen. The data available here included data on the level of individuals from the socioeconomic census of 2001 conducted in Belgium and marginal data available for the variables of interest that were desired to be controlled for, for the Belgian population in the year 2010. At the household level, the variables controlled for included: availability of cars in a household, age of the householder and household size. At the individual level, gender and age were controlled for. To estimate the target joint distributions for Flanders in 2010, the socioeconomic census joint distributions were updated using the Iterative Proportional Fitting (IPF) algorithm based on the population marginal of 2010 for Belgium. This was conducted both at the household and the person level based on the control variables mentioned above. The results obtained from comparing the generated synthetic populations with the real data provided support that both the household and the person level distributions of the control and some non-control variables represented the true population well and consequently the actual population could be relatively accurately synthesized.

Environment data

a) Zoning system

The unit of geography inside the FEATHERS framework is defined by means of a hierarchy of three geographical layers on top of each other. This hierarchy stems from the land use data being available at different levels of geographical detail.

In order of increasing detail a list of Superzones, Zones and Subzones were developed for the 2010-based FEATHERS framework. The list of Superzones corresponded with all municipalities inside Belgium, the Zones corresponded with administrative units at one level lower than municipalities and the last level, the Subzone level, consisted of virtual areas that were constructed based on homogeneous characteristics defining each Subzone.

b) Land use data
Data about land use were available at different levels of the zoning system and involved opening hours and locations of facilities for out-of-home activities. Moreover, the land use system also provided sector-specific data that were used as indicators of availability and attractiveness of facilities for conducting particular activities. In the FEATHERS database the following sector-specific data are being used: total employment, number of primary school children, employment in daily good retailing, employment in non-daily good retailing, employment in banks/post offices and employment in restaurants. By means of assumed relationships between these sector data sets and activity types such as 'shopping', 'bring/get', etc., possible locations for conducting such kind of activities could be obtained.

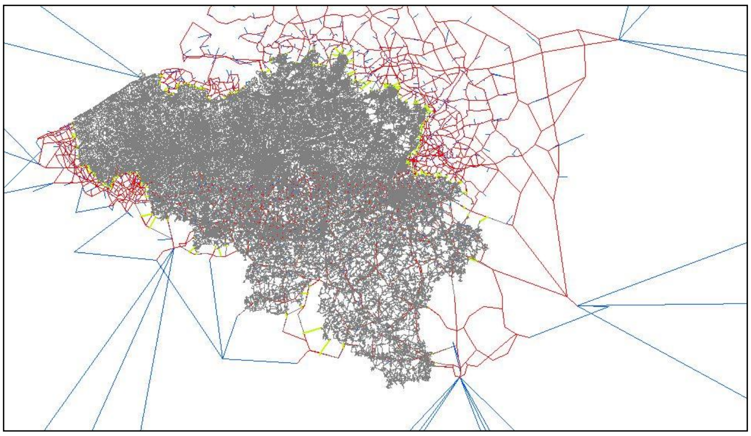
3) Widening and incorporation of supply data (networks, roads, intersections, …) for the full region of Belgium in the FEATHERS framework

In order to supply FEATHERS with the necessary network information and in order to perform traffic assignments, a network for the Belgian situation had to be build. This task was performed in the geographical information system package TransCAD.

A network is a special data structure that stores important characteristics of transportation systems and facilities. A typical TransCAD network is defined, derived, and used in conjunction with a line layer and its associated endpoint layer. The network that was created for this study contained all the origin and destination nodes that correspond with the Subzone layer. Even though for this study it was possible to start with an already existing Belgian network, still some manipulations had to be done in order to obtain a fully operational network that could be used for FEATHERS. One of the important steps was to create for each Subzone a list of centroid connectors. Centroid connectors are not physical links but are necessary in order to derive information between each origin-destination combination. This derivation was done by means of a special technique. When shorthest paths are generated, one might also want to compute the total value of other network attributes along the path. The technique of computing these secondary attributes is called skimming. By skimming the modified Belgian network, data about for example Free floating car travel time, could be derived. After skimming the Belgian network, the data was stored into FEATHERS.

The transportation system in FEATHERS is represented by a set of Level Of Service (LOS) matrices by transport mode containing information about travel distance, travel time, egress and access time. The different transport modes that were considered included, car (driver), car (passenger), public transport and slow mode. Each transport mode had its own pair of distance and travel time matrices. These travel distances and travel times were derived in a pre-processing stage outside the system using TransCAD as described above. Furthermore, in the FEATHERS transport demand model it was also important to have time-of-day dependent travel times rather then an average. For this reason, free-flow, morning-peak and evening-peak travel time data were derived accordingly.

For the purpose of an illustration, the picture below gives an impression of how the modified Belgian network looks like.



4) Creation of a data imputation method with support vector machines for activity-based transportation models

Activity-based approaches in transportation models aim at predicting which activities are conducted where, when, for how long, with whom, the transport mode involved and so on. An activity-based framework named FEATHERS for Flanders in Belgium has been developed at IMOB since 2005. During the establishment of the framework, lots of data are needed. One of the main data sources are activity-based diaries. However, activity diaries tend to contain incomplete information due to various reasons. More recently, with the development of computer science and technology, some artificial intelligence and machine learning techniques have arisen to process the missing data. In this research, a data imputation method with a Support Vector Machine (SVM) is proposed to solve the issue of missing data in activity-based diaries. In order to verify the efficiency of SVMs, other methods such as LDA (Linear Discriminant Analysis) and PNN (Probabilistic Neural Network) were also used to process the same data imputation problem. Compared with accuracies obtained by SVMs, the accuracies obtained by LDA and PNN proved to be lower. The initial experimental results showed that missing elements of observed activity diaries could be accurately inferred by relating different pieces of information. Therefore the proposed SVM data imputation method in this research served as an effective data imputation method that can induce complete activity diaries in the case of missing information.

The activity-based approach is a sound option to model people's travel behavior, which has set the standard for travel demand modeling during the last decade. The basic premise of this approach is that travel demand is derived from the activities that individuals and households need or wish to perform. A dynamic activity-based travel demand framework, FEATHERS has been developed for Flanders (the Dutch speaking region of Belgium) based on the above aim. The FEATHERS framework to be applied for the whole Belgium has also been developed at IMOB during this research period. To build the FEATHERS model that can predict all of those above facets, one requires data on all these facets. Clearly, the data collection is a huge challenge. One of the main data sources is activity-based diaries. These diaries are considered to be the most important source of information that benefits the establishment of transportation models. However, activity-based diaries have also been proven to have many disadvantage. One is that the diaries demand high effort to plan and implement and also require high costs in terms of time, finances and other resources. The other is that the collection of diary data frequently brings along a huge burden on respondents to maintain and recall exact details. Consequently, activity diaries tend to contain incomplete information due to various reasons, which is a serious problem because activity-based models require complete diary information.

Activity diaries used in the existing FEATHERS contain a combination of individual surveys and household surveys. Each household or person sample includes many variables. Among all samples, about 10% households samples and 5% person samples contain missing information. The missing information in a sample may contain one missing variable value, two missing variable values or more than two. If all samples that contain any missing values are deleted and the analysis is then carried out on the samples that remain, some serious drawbacks will be brought. One of the drawbacks is the reduction of samples, which will affect the predicting reliability and quality of the FEATHERS model. The other is that the elimination of useful information in the sample will result in serious biases if the samples are not missing completely. The interest of this research has centered on performing data imputation, the process by which missing values in a data set are estimated by appropriately computed values, thus constructing a complete data set.

More recently, with the development of computer science and technology, some artificial intelligence and machine learning techniques have arisen in the area of missing data treatment, such as artificial neural networks (ANN), fuzzy logic systems, and rough sets, which stimulate the missing data research to a new stage. An ANN is a mathematical model or computational model that is inspired by the structure and/or functional aspects of biological neural networks. Modern neural networks are usually used to model complex relationships between inputs and outputs or to find patterns in data. However, the problem of over-training has emerged in neural networks, which results in a low generalization. The support vector machine (SVM) is a new generation of learning systems based on recent advances in statistical learning theory. The advantage of using SVM over other methods is its high quality of generalization. SVMs have been applied in many areas such as text categorization, hand-written character recognition, image classification, and bio-sequences analysis. The very first introduction of SVMs in the early 1990s lead to a recent explosion of applications and deepening theoretical analysis, which has now established SVMs for machine learning and data mining. In this research, the SVM was proposed to predict the missing values of two variables. Here two SVM models were established to predict the missing elements of 'number of car' and 'driver license' respectively. The first SVM model to predict the number of car achieved an accuracy of 69%. Meanwhile, the second SVM model to predict driver license could obtain an accuracy of 83 %. The results were verified by a four-fold cross-validation. Then, the SVM method was compared with the Probabilistic Neural Network (PNN) Algorithm and linear discriminant analysis (LDA). The results showed that the SVM can obtain a higher accuracy than the PNN and LDA. The SVM provides a feasible solution to the missing information in the FEATHERS model, by which the missing data were estimated and complemented. Available data could be utilized by the greatest extent and so the reliability of the FEATHERS model could be improved.

Since the objective of the activity diary is to give a representative description of the travel behavior of the population in Belgium, the target population in the project was defined as "all the people residing in Belgium, regardless of their place of birth, nationality of any other characteristics". Activity diaries used in FEATHERS contain a combination of individual surveys and household surveys. The data were collected in 2010, in which the individual surveys were carried out among Belgian citizens aged 6 years and above. The total number of collected samples equals 8551 households comprising 15888 individuals. In the activity diary, a household record has many variables, such as 1) the name, gender, nationality, educational certificate, and professional status of each household member; 2) the type of vehicle, number of the specified vehicle, and purchase year of the specified vehicle that the household possess; 3) the place of residence, net income of the household, etc. The individual survey includes person ID, the mode of travel, number of trips, the start time, the arrival time, activity type, activity duration, activity location ID, and the driver license. All data had to be preprocessed to meet the requirements of FEATHERS. which need five input files (Households, Persons, Activities, Journeys and Lags). Each file includes many variables respectively. Here the interest centered on the Households and Persons. There are eight variables (HouseholdID, Household locationID, Household composition, Socio-economic class, Age oldest household member, Children age class, Number of cars and Number of household members) in the Households file. Meanwhile, the Persons file includes six variables (PersonID, HouseholdID, Personage, Work status, gender and Driver's license). Among all samples, about 10% of household samples and 5% of person samples contained missing information. How to estimate the missing information and so improve the number and quality of samples has been the main concern of this research. The following section will describe a SVM method to process the missing information.

A SVM is one of the supervised learning methods that analyze data and recognize patterns, used for classification and regression analysis. It takes a set of input data and predicts which of two possible classes each given input belongs. The SVM performs classification by constructing a hyperplane that optimally separates the data into two categories. New examples are then mapped into that same space and predicted to belong to a category based on which side of the hyperplane they fall on. A good separation is achieved by the hyperplane that has the largest distance to the nearest training data points of any class (so-called margin), since in general the larger the margin the lower the generalization error of the classifier. So the key technique to SVMs is how to obtain a hyperplane that has the largest margin. In general, an original problem may be stated in a finite dimensional space. The maximum-margin hyperplane needs to be obtained by transforming the original space to a high-dimensional space, which is proposed to make the separation easier in that space. To keep the computational load reasonable, the SVM scheme is designed to ensure that it can be computed easily by a specific function in the original space. This problem can now be solved by standard quadratic programming techniques and programs.

As described earlier, there are eight variables that are needed in the Households file to meet the requirements of FEATHERS. However, about 8% of the samples in this file miss information on 'number of cars'. Since different variables are interrelated, the missing variable information in one sample is expected to be inferred by other variables in the same sample. Here the missing 'number of cars' are predicted by other five variables. The inputs of the SVM model include Household composition, household income, Age oldest household member, Children age class and Number of household members. The output of the SVM is the 'number of cars', which has three values (0, 1, 2). These three values represent that a household has no car, one car, two cars or more, respectively. Obviously, the prediction is a three-category classification. However, a SVM is often used to distinguish two categories. The approach used in this research for doing so was to reduce the single multiclass problem into multiple binary classification problems. First, a strategy called "one against many" was adopted, in which each category was split out and all of the other categories were merged. According to this strategy, three SVM models (SVM Model 1, SVM Model 2, and SVM Model 3) were established. The SVM model 1 was used to distinguish label 0 and the rest, where label 1 and label 2 were merged. The SVM model 2 was used to distinguish label 1 and the rest, where label 0 and label 2 were merged. The SVM model 3 was used to distinguish label 2 and the rest, where label 0 and label 1 were merged. The classification accuracy was obtained by means of a four-fold cross-validation. By comparing results, the SVM model 1 obtained the highest classification accuracy. And so the SVM model 1 was selected to distinguish label 0 and the rest. If a new sample would be assigned as label 0, the classification would be over. Otherwise, the sample would need to be distinguished continuously between 1 and 2. After the classification of the SVM model 1 and the SVM model 4, the final label was decided. The SVM model to predict the 'number of cars' achieved an accuracy of 69% by means of a four-fold cross-validation.

As described earlier, the Persons file used in FEATHERS include six variables (PersonID, HouseholdID, Personage, Work status, gender and Driver's license). However, about 2% samples in this file was missing information of 'driver license'. Considering the relativity between Person age, Work status, gender and driver license, here the missing 'driver license' data were predicted by the other three variables (Person age, Work status, and gender). Therefore, the inputs of the SVM model included Person age, Work status and gender. The output of the SVM was the 'driver license', which had two values (0 and 1). These two values represent that a person has no driving license or has a driving license. The SVM model to predict the presence of the driving license obtained an accuracy of 83 % by means of a four-fold cross-validation.

Neural networks are frequently employed to classify patterns based on learning from training samples. Different neural network paradigms employ different learning rules, but all in some way determine pattern statistics from a set of training samples and then classify a new unknown sample on the basis of these statistics. A PNN is a useful neural network architecture and includes an input layer, a hidden layer and an output layer. The PNN is a supervised learning algorithm but includes no weights in its hidden layer. Instead each hidden node represents a training vector, with the training vector acting as the weights to that hidden node. The input layer represents the feature vector of a new sample, which is fully interconnected with the hidden layer. The actual training vector serves as the weights as applied to the input layer. Finally, an output layer represents each of the possible classes for which the new sample can be classified. The output class node with the largest activation represents the winning class. In this research, the PNN and the LDA were also used to predict the 'number of cars' and the existing of 'driver license'. The input of the PNN and the LDA were the same as the one of SVM.

The following section discusses the results. In this research, two SVM models were established to predict the 'number of cars' and 'driver license' respectively. During the prediction for ‘number of cars’, two necessary steps were taken. In the first step, a strategy called “one against many” was adopted and the best result was obtained by distinguishing between label 0 and the rest. The label 0 represented a household that did not have a car and the rest (label 1 and label 2) represented a household that did have one car, two cars or more. The obtained result showed that a household without any car can be best differentiated from the one with cars, which is concordant with our intuitive impression. In our real life, a household without any car has many differences with the one with cars according to our feeling. We also established models for any two categories, that was between label 0 and label 1, label 0 and label 2, label 1 and label 2. The classification results of these two categories were 78%, 90% and 79% respectively, which showed that a household without any car can be distinguished from the one with two cars better than other two instances. All these results showed that the established models were consistent with real life. If a new sample would be assigned ‘the rest’ in the first step, it would then be assigned to either label 1 or label 2 in the second step. In the prediction of ‘number of cars’ and ‘driver license’, the accuracies of 69% and 83% were obtained respectively by means of a four-fold cross-validation. The missing information in these two variables was imputed by related known information in other variables. The prediction results kept identical with analyzed facts, which showed that the established SVM models were feasible to complete information in the FEATHERS model. In order to verify the efficiency of the proposed SVMs, the PNN and the LDA were also used to predict the ‘number of cars’ and ‘driver license’. However it could be shown that the SVM was the most superior data imputation method.

This last section covers the conclusions that can be made. The establishment of the FEATHERS model needs large amount of complete data. How to improve the quality and quantity of sample data under existing activity-based diaries was one of problems to be solved in FEATHERS. Aiming at this problem, a data imputation method based on SVM was proposed. Two SVM models were established to predict the missing information of variables called ‘number of cars’ and ‘driving license’ using related other variables. The prediction accuracies of 69% and 83% were obtained respectively by means of a four-fold cross-validation. After comparing with the LDA and PNN methods, the initial results showed the feasibility of the proposed method. The complete sample number could be increased by 1%-8% after an accurate prediction. The established SVM models can now be used to new samples, which provide a good approach to perfect the missing data.

Publications resulting from the work done

Banghua Yang, Davy Janssens, Da Ruan, Tom Bellemans, and Geert Wets, A Data Imputation Method with Support Vector Machines for Activity-Based Transportation Models. Book chapter forthcoming in *Computational Intelligence for Traffic and Mobility*