Available online at www.sciencedirect.com
ScienceDirect
Procedia Computer Science 1 (2012) 1259–1265
Procedia Computer Science
www.elsevier.com/locate/procedia

International Conference on Computational Science, ICCS 2010

Towards automated model calibration and validation in rail transit simulation

Yilin Huang, Mamadou D. Seck, Alexander Verbraeck

Systems Engineering Group, Faculty of Technology, Policy and Management
Delft University of Technology, The Netherlands

Abstract

The benefit of modeling and simulation in rail transit operations has been demonstrated in various studies. However, the complex dynamics involved and the ever-changing environment in which rail systems evolve expose the limits of classical simulation. Changing environmental conditions and second order dynamics challenge the validity of the models and seriously reduce model (re-)usability. This paper discusses the potential benefits and requirements of dynamic data-driven simulation in rail systems. The emphasis is placed on automated model reconfiguration, calibration, and validation through the use of data analysis methods. The rationale and requirements are discussed and a process model for data driven calibration and validation is proposed.

Keywords: Public rail transport simulation, model validation, calibration

1. Introduction

Rail transport systems have inherent long life spans. In addition to their well-known dynamic characteristics known as "first order dynamics", the system themselves are also subject to change, which is referred to as their "second order dynamics" [1]. In public rail transport systems, daily/weekly/seasonal patterns of vehicle ridership, and the driving/speed profile of a vehicle are examples of first order dynamics; second order dynamics often denotes the maintenance and evolution of the system [1], e.g., adjustment of timetables, alteration of control strategies, and extension of rail networks, which could further cause the change in first order.

In simulation studies, relevant and correct data is needed to formulate a model and, if pertinent, the environment that provides the input to excite the model and to affect the model behavior [2]. The leaping advances in data collection and storage technology in the recent years have enabled organizations to accumulate vast amount of data [3]. In modern rail transit operation, many vehicles possess Automated Vehicle Location (AVL) and Automatic Passenger Counter (APC) devices that are able to collect the data. These data provide us with a rich source to study and trace the dynamic features within the system. Dynamically incorporating such type of data into a simulation could offer

Email addresses: y.huang@tudelft.nl (Yilin Huang), m.d.seck@tudelft.nl (Mamadou D. Seck), a.verbraeck@tudelft.nl (Alexander Verbraeck)
URL: http://simulation.tudelft.nl, www.tudelft.nl/yhuang (Yilin Huang), www.tudelft.nl/mseck (Mamadou D. Seck), www.tudelft.nl/averbraeck (Alexander Verbraeck)

1877-0509 © 2012 Published by Elsevier Ltd.
doi:10.1016/j.procs.2010.04.140

more accurate predictions and more reliable outcomes [4]. As data are processed, simulation models are adjusted to best agree with the measurements [5]. Enabling the capability entails extra functionality into the simulation model, which inevitably introduces greater complexity; however, for long life cycle models [6] of complex systems, such an approach is still beneficial, especially if part of the model components could be designed as generic as possible.

The traditional approach for modeling captures first order dynamics in which the predefined input data forms the boundary and the environment of the model. Such an approach falls short of tracing second order dynamics as no extra (new) data is being fed to the model. Consequently, changing a model built along a traditional approach to incorporate the second order dynamics of the system will cause extensive model reconstruction/modification, calibration, and validation. These activities are often labor-intensive, time-consuming, and error-prone.

This paper looks into the use of a dynamic data-driven approach for rail transport simulation that aims at capturing both first and second order dynamics of the system from the onset. The main effort at this stage is to employ computational logic that continually performs model validation and calibration using the extensive available data during a simulation run. The remainder of this paper is organized as follows. In the next section, the background of this research is presented. Some related research is reviewed in section 3. It is followed by the challenges of data analysis in dynamic data-driven rail transit simulation. Section 5 discusses the proposed data-driven simulation process.

2. Background

A simulation model may have a short or long life cycle, depending on the use of the model through the life of the real system it represents [6]. Short-life-cycle models are for a single or only few decision making purposes. Once the decision is made, the model is discarded. Long-life-cycle models are used at multiple points in time during the life of the real system and are maintained and revalidated as conditions change in the system. Such models are generally built for the purpose of design, operation, training, etc. of the system [6, 7, 8]. Rail transport systems are typical examples of the second case. They are often complex, large-scale, and require huge investments and long term development. Therefore there is much interest in developing reusable rail models. In this context, the dynamic data-driven approach could offer useful functionalities such as detecting situation or scenario changes and the corresponding system behavior changes, and performing automated model or sub-model adaptations adequate to the changes. Herein lies the important roles of automated model validation and calibration.

The dynamic data-driven approach for rail transport simulation is conceived in line with the DDDAS (Dynamic Data Driven Application Systems) concept [9]. DDDAS entails the ability to dynamically incorporate additional data into an executing application, and in reverse, the ability of an application to dynamically steer the measurement process [10]. A similar concept is addressed by the term Symbiotic Simulation Systems [11], which emphasizes a close relationship between a simulation system and a physical system that is mutually beneficial. The physical system benefits from the optimized performance which is obtained from the analysis of the simulation experiments, while the simulation system benefits from the continuous supply of the latest data [12]. On-line simulation is probably the oldest term which has been used for the concept of coordinating simulation with the dynamic evolvement of the real system being studied. However this term is used in various ways, and its definition is rather vague [13, 14, 15]. This presumably caused simulation practitioners to coin more accurate terms and definitions.

At the moment of writing this paper, a few projects have been carried out by the authors' research group concerning simulation of rail transport systems. A library of rail simulation components has been developed [16, 17]. It is an open source Java package that supports distributed microscopic multi-formalism simulation of heavy and light rail operations, which helps rail designers to analyze the (first order) dynamics in rail transport systems, and to assess the impact of the network design including control strategies and timetables. Several case studies show that the tool is helpful to improve the reliability of the rail design [18, 19, 20]. The library provides a foundation for dynamic data-driven simulation of rail transport systems. Components that perform data processing, analysis, and parameter estimation will be added to the library as an extension to support dynamic data-driven simulation.

3. Related Work

In recent years, more research efforts are devoted to the efficient use of data in transport simulation in order to better cope with the intrinsic dynamic nature of the system. Some emerging simulation-based tools are presented below.

The automated signal timing plan procedure in [21] uses archived traffic data. It applies hierarchical cluster method to identify time intervals in which traffic patterns are relatively constant and to determine the traffic volumes in each interval.

The traffic signal control system *HUTSIG* is based on on-line simulation that connects to real-time data [22]. Each signal is an agent that negotiates with other agents about the control strategy. A fuzzy inference engine resembles human (expert) reasoning processes.

An on-line traffic management support tool [23] compares real-time traffic data with the achieved data. If a problem pattern is identified, several predefined countermeasures will be proposed, and simulation will help to choose the most adequate option.

The real-time freeway network surveillance tool described in [24] is based on a macroscopic traffic flow model. This on-line tool estimates traffic flow variables and model parameters over a short-term time horizon. The methods applied include extrapolation of historical data, dynamic traffic flow modeling, and spatial interpolation of detector measurements. The adaptive features of the traffic state estimator (the performance of on-line model calibration, auto-adaptation to external conditions, and enabling of incident alarms) are further evaluated in [25].

Regiolab Delft traffic laboratory [26] collects live traffic data from various sources, e.g., electronic measuring systems on highways, inductive loop detectors at intersections, license plate readers, infra-red measuring systems on rural roads. In a simulation study [27], drivers are agent-based self-learning driving behavior models, and data is fed to a belief network coupled with a decision network to "train" the agent to predict the future traffic situation evolvement with a rolling horizon of 30 minutes.

The research project at *Georgia Tech* concerns data-driven simulation of surface transportation systems [28, 29, 30]. The research uses real-time aggregated data streams to estimate the evolving state of traffic. Data streams with various update intervals were examined to study which is the best to drive the simulation for an appropriate representation of the actual traffic situation. They also discussed the concept of ad hoc distributed simulations, where individual in-vehicle simulation captures the "near-by" measurement data and models a portion of the traffic network, from which the server constructs an overall picture of the entire network [31]. Some experiment results of the distributed simulation implemented by a cellular automata model and by a *VISSIM* model are presented in [32].

4. Challenges of Data Analysis in Dynamic Data-driven Rail Transit Simulation

Systems characterized by nonlinear interactions often exhibit emergent behaviors that are difficult to predict based on their initial conditions; hence traditional simulations that rely on such conditions are likely inaccurate or invalid in simulating these systems [33]. Furthermore, because changes in second order are hard to predict, and they also affect changes in first order, the distinction between first and second order dynamics in systems and their representations is not generally recognized or modeled [1]. In practice, if second order dynamics occurred, the model would be modified and revalidated. But with the dynamic data-driven concept, modeling second order dynamics becomes possible.

The authors discussed the dynamic data-driven simulation approach, rail operation data classification, and some data analysis methods in [34]. The simulation approach (figure 1) continually compares the model output and system measurement. If the discrepancy exceeds a predefined threshold, the model parameters that are observable from the system are updated, and more importantly the unobservable parameters are estimated based on the current situation and historical data. (The data are classified, e.g., per time, station, and distance, in the data pre-processing phase.) The "learning phase" continues with the hope that the modeled result will have a convergence with the measurement from the real system. This model can be hereafter used to simulate faster than real-time giving in-time system predictions.

The research aims at enhancing dynamic data-driven simulation with a focus on model calibration and validation through on-line data analysis. Such simulation will typically require integrated data processing and analysis mechanisms that execute in parallel with the simulation. Deriving model parameters from measurement data necessitates in-depth data analysis. In case of the train, light rail or tramway models of a complex transport network, there can be dozens of lines, hundreds of stops or stations, hundreds to a few thousand of individual distances between stops or crossings, each with AVL and APC data for more than a hundred vehicles per day (daily pattern), with differences for the weekdays (weekly pattern). Analysis of a month of data can easily lead to a data set with tens to hundreds of millions of data points to help in the estimation of the thousands of parameters in the model.

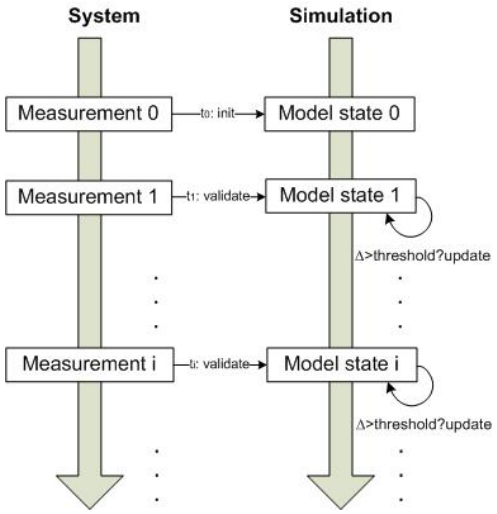

Figure 1: The simulation approach

Data classification is important towards tackling the data analysis issue, because data in different categories may ask for different data analysis methods. These methods are necessary for different tasks, such as to study the distribution and spread of data, to prepare for hypotheses test, and to fit distributions for the model [35]. For example, proximity-based techniques or density-based techniques [3] can be used to identify outliers where extreme long travel times shall be discarded when typical travel times are to be determined; periodogram analysis [36] may be used to study daily (based on time-of-day) and weekly (based on day-of-week) patterns of travel behavior; interpolation and extrapolation can be used on AVL data sets to estimate the vehicle location at a designated time point.

Many methods and algorithms exist for the analysis of data and time series. For parameter estimation, a number of steps have to be taken, such as cleaning the data for outliers by either leaving out individual data points or complete data sets, separating the "normal" behavior in the system from "disturbed" behavior or "changed" behavior; some examples are shown in [34]. The difference between outliers and behavior to be modeled is difficult to make. When calibrating the model parameters according to the data feed, it is important to know whether the deviation of the expected behavior (or model output) is due to incorrect model configuration or due to behavior changes of the system. Such judgements can be rather subjective, and may change due to circumstances. It is already difficult for the human analysts, and even more difficult for the computer. However, given the amount of data for models like those of train or light rail company, automated procedures are needed. Such procedures shall be able to populate the models with valid data, compute correct interpretation, and find useful data patterns for parameter estimation.

5. The Dynamic Data-Driven Simulation Process

In this section, the simulation process is introduced. The process is illustrated in figure 2. The goal is to design the components as generic as possible, so that they may be jointly used with other tools in other applications. First of all, depending on the parameters to be calibrated, data sets need to be classified and prepared for analysis. This can cover activities such as data (table, record and attribute) selection, cleaning and transformation. For example, to determine the distribution and parameters of dwell times at a stop according to time-of-day, the daily pattern of the dwell time will be of interest and the corresponding data will be selected. Some detailed issues concerning data preparation shall be be addressed: e.g., whether to use sampling or aggregation in case the data set is too large (which is often the case); where to put the threshold to rule out outliers; if outliers shall be analyzed in a separated case, and be generated in the simulation; whether to separate peak and off-peak measurements or weight them accordingly for statistics; or if some

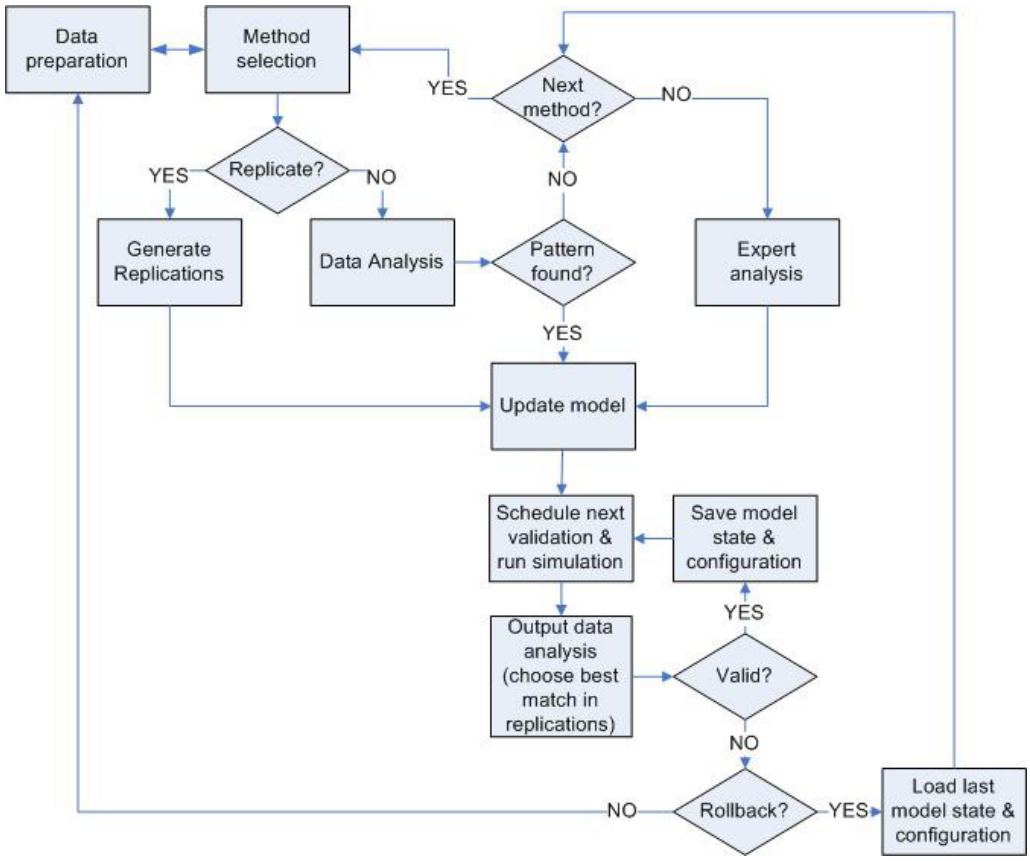

Figure 2: The simualtion process

data are missing, how to e.g., insert suitable default values. These and many other issues need to be premeditated with the goal to facilitate and ease the data analysis step.

With the data sets correctly selected and transformed, the next step is to choose and apply the analysis methods. The appropriate methods or algorithms for certain data sets are predetermined during the design phase to suit best the characteristics of the data sets and of the relationship between the data sets and the desired model parameters. Several methods can defined for parameter fitting or estimation. The methods can be applied successively if the result of a previous one doesn't satisfy the validation step. Note that a newly selected method may require different or extra data preparation. In case the specific methods are not sufficient in finding satisfactory parameter estimation, a stochastic approach can be taken. Experiments with random inputs or input combinations can be run with several model replications. (Later by comparing the response variances, a best match can be chosen.)

The selected method needs to run on the data set to reveal hidden patterns. Success criteria shall be computed to assess if a useful pattern is found. The concern arose because if a data set is searched long enough, one can always find patterns that appear to be statistically significant but, in fact, are not [37]. If the result is positive, the discovered pattern will be used for parameter update. If the pattern is not useful (e.g., the same pattern is found before for similar situations but the validation disproved the parameter estimation), another analysis method can be applied. If no methods (including the stochastic approach) succeed to find a useful parameter estimation, expert interference can be supported by interactive processes, which would make the system fall back to the traditional way of readying the data for a simulation study.

Once the model is updated, simulation is run to validate the parameter estimation. The simulation run length is application dependent. It can be long, e.g., in hours, if one wants to study the effect of an accident, and how the traffic would recover from it. The run length can also be short, e.g., in minutes, if one wants to trace and simulate individual vehicles, detect instant changes and abnormalities, and assess the corresponding control strategies and solutions.

The validation performed is rather a comparison. It is expected that the simulation output and the relevant measurement from the system shall reflect the same phenomena (thus converge) if the simulation has the same initial state as the system and the model is a valid one. Given that the conceptual model has been beforehand validated, the correctness of the parameter configuration will determine the convergence of the simulation output and measurement. It is hence theoretically not hard to find out if the parameter configuration is valid. A number of details have to be planned, e.g., what key performance indicators will be compared. In this context, statistical methods are again useful to extract quantitative descriptions.

If the model has passed the validation, the current model state and parameter values can be saved with a time stamp for rollback or future reference, and the simulation can be continued for the next round of validation. The purpose of automated calibration and validation is to detect unanticipated changes and events and to adapt the model to such randomness. Thus a valid model configuration at an instance may not be valid at a next moment. Once it is detected to be invalid, a new iteration of parameter calibration will be performed. The model can be rolled back to a previous valid state, if the same data set (but with updated data values) shall be analyzed. In this case, another method can be chosen. If different data sets shall be chosen, e.g., when the same data set has been analyzed but the simulation never passed the validation, the rollback can be skipped, and other criteria can be used to select data sets and start a next iteration of the process.

Another point needs to be mentioned is that independent data sets shall be used for simulation experiments and validation. In the experiments, the automated parameter calibration and validation process are studied and tested. Validation of the process itself thus needs completely different data sets. For example, one can use measurement data collected in the odd weeks for experiments, and data from even weeks for validation.

6. Conclusions

In this paper, we have discussed dynamic data-driven model update, calibration and validation in the context of rail transit simulation. Different challenges have been identified such as handling large volume of data, classification of data, the selection of data, the implementation of different data analysis methods, and the difficulties of data interpretation. The proposed process is a step towards a generic library for dynamic data-driven model calibration and validation. In this process, the model output and measurements from the system are continually compared. Once the deviation exceeds a predefined threshold, the model parameters should be updated based on data analysis results. Different analysis methods need be available to discover useful hidden patterns in the data sets, and to estimate the parameters. Mechanisms should be defined to enable saving model states and parameter configurations as a reference for ulterior comparison or model state rollback. The proposed process will be implemented as an extension of the open source Java rail simulation library [16, 17]. The performance assessment of the components will be carried out in case studies of light rail projects of HTM, a public transport company in The Hague, the Netherlands.

7. Acknowledgments

The authors would like to acknowledge the support of HTM Urban Public Transport, The Hague, The Netherlands.

8. References

[1] G. J. Ramackers, A. A. Verrijn-Stuart, First and second order dynamics in information systems, in: H. G. Sol, K. M. v. Hee (Eds.), International Working Conference on Dynamic Modelling of Information Systems, 1991.

[2] T. I. ren, Impact of data on simulation: From early practices to federated and agent-directed simulations, in: A. H. et al. (Ed.), Proceedings of the EUROSIM 2001, 4th International Eurosim Congress, 2001, pp. 3–8.

[3] P. N. Tan, M. Steinbach, V. Kumer, Introduction to Data Mining, Addison-Wesley, 2005.

[4] DDDAS, Dynamic Data Driven Application Systems, Workshop report, NSF DDDAS 2006 Workshop, `http://www.nsf.gov/cise/cns/dddas`, Virginia, USA (January 2006).

[5] C. C. Douglas, Dynamic Data Driven Applications Systems - DDDAS 2008, in: ICCS 2008, Part III, LNCS, Vol. 5103, Springer-Verlag, Berlin, 2008, pp. 3–4.
[6] O. Ulgan, A. Gunal, Simulation in the automobile industy, in: J. Banks (Ed.), Handbook of Simulation: Principles, Methodology, Advances, Applications, and Practice, Wiley Interscience, 1998, pp. 547–570.
[7] Y. A. Saanen, A. Verbraeck, J. C. Rijsenbrij, The application of advanced simulations for the engineering of logistic control systems, in: Proceedings ASIM 2000 - The new simulation in production and logistics, 2000, pp. 215–229.
[8] C. Versteegt, A. Verbraeck, The extended use of simulation in evaluating real-time control systems of agvs and automated material handling systems, in: Proceedings of the 2002 Winter Simulation Conference, 2002, pp. 1659–1666.
[9] DDDAS, Creating a dynamic and symbiotic coupling of application/simulations with measurements/experiments, Workshop report, NSF DDDAS 2000 Workshop, `http://www.nsf.gov/cise/cns/dddas`, Virginia, USA (March 2000).
[10] F. Darema, Dynamic Data Driven Application Systems: A New Paradigm for Application Simulations and Measurements, in: M. B. et al. (Ed.), ICCS 2004, LNCS, Vol. 3038, Springer-Verlag, Berlin, 2004, pp. 662–669.
[11] R. Fujimoto, D. Lunceford, E. Page, A. Uhrmacher, Grand challenges for modelling and simulation, Tech. rep., Dagstuhl-Seminar (2002).
[12] H. Aydt, S. J. Turner, W. Cai, M. Y. H. Low, Symbiotic simulation systems: An extended definition motivated by symbiosis in biology, in: Proceedings of 22nd Workshop on Principles of Advanced and Distributed Simulation, 2008, pp. 109–116.
[13] M. M. Jones, On-line simulation, in: Proceedings of the 1967 22nd ACM Annual Conference/Annual Meeting, 1967, pp. 591–599.
[14] W. J. Davis, On-line simulation: need and evolving research requirements, in: J. Banks (Ed.), Handbook of Simulation: Principles, Methodology, Advances, Applications, and Practice, Wiley Interscience, 1998, pp. 335–393.
[15] F. Kamrani, R. Ayani, Using on-line simulation for adaptive path planning of UAVs, in: Proceedings of the 11th IEEE International Symposium on Distributed Simulation and Real-Time Applications, 2007, pp. 167–174.
[16] E. M. Kanacilo, A. Verbraeck, A distributed multi-formalism simulation to support rail infrastructure control design, in: Proceedings of the 2005 Winter Simulation Conference, IEEE, 2005, pp. 2546–2553.
[17] E. M. Kanacilo, A. Verbraeck, Simulation services to support the control design of rail infrastructures, in: Proceedings of the 2006 Winter Simulation Conference, IEEE, 2006, pp. 1372–1379.
[18] E. M. Kanacilo, A. Verbraeck, Assessing tram schedules using a library of simulation components, in: Proceedings of the 2007 Winter Simulation Conference, IEEE, 2007, pp. 1878–1886.
[19] E. M. Kanacilo, N. v. Oort, Using a rail simulation library to assess impacts of transit network planning on operational quality, in: WIT Transactions on the Built Environment, no. 103, WIT Press, Southampton, 2008, pp. 35–43.
[20] Y. Huang, A. Verbraeck, N. van Oort, H. Veldhoen, Rail transit network design supported by an open source simulation library: Towards reliability improvement, in: Transportation Research Board 89th Annual Meeting Compendium of Papers, no. 10-0310, 2010.
[21] B. L. Smith, W. T. Scherer, T. A. Hauser, B. B. Park, Data-driven methodology for signal timing plan development: A computational approach, Computer-Aided Civil and Infrastructure Engineering 17 (6) (2002) 387–395.
[22] I. Kosonen, Multi-agent fuzzy signal control based on real-time simulation, Transportation Research Part C: Emerging Technologies 11 (5) (2003) 389–402.
[23] J. Barcel, M. Delgado, G. Funes, D. Garca, A. Torday, On-line microscopic traffic simulation to support real time traffic management strategies, in: Proceedings of the 6th ITS European Congress, no. 2607, 2007.
[24] Y. Wang, M. Papageorgiou, A. Messmer, A real-time freeway network traffic surveillance tool, IEEE Transactions on Control Systems Technology 14 (1) (2006) 18–32.
[25] Y. Wang, M. Papageorgiou, A. Messmer, P. Coppola, A. Tzimitsi, A. Nuzzolo, An adaptive freeway traffic state estimator, Automatica 45 (2009) 10–24.
[26] T. H. J. Muller, M. Miska, H. J. v. Zuylen, Monitoring traffic under congestion: Base for dynamic assignment in online prediction models, in: 84th Transportation Research Board 2005 Annual Meeting, 2005.
[27] M. P. Miska, T. H. J. Muller, H. J. v. Zuylen, Calibrating driving behavior with microscopic measurement data, in: Proceedings of the 2006 IEEE Conference on Intelligent Transportation Systems, 2006, pp. 1609–1614.
[28] M. Hunter, R. Fujimoto, W. Suh, H. K. Kim, An investigation of real-time dynamic data driven transportation simulation, in: Proceedings of the 2006 Winter Simulation Conference, 2006, pp. 1414–1421.
[29] R. Fujimoto, R. Guensler, M. Hunter, H. Kim, J. Lee, J. Leonard II, M. Palekar, K. Schwan, B. Seshasayee, Dynamic data driven application simulation of surface transportation systems, in: ICCS 2004, Part III, LNCS, Vol. 3993, Springer-Verlag, Berlin, 2006, pp. 425–432.
[30] D. Henclewood, M. P. Hunter, R. M. Fujimoto, Proposed methodology for a data-driven simulation for estimating performance measures along signalized arterials in real-time, in: Proceedings of the 2008 Winter Simulation Conference, 2008, pp. 2761–2768.
[31] R. Fujimoto, R. Guensler, M. Hunter, K. Schwan, H. Kim, B. Seshasayee, J. Sirichoke, W. Suh, Ad hoc distributed simulation of surface transportation systems, in: ICCS 2007, Part I, LNCS, Vol. 4487, Springer-Verlag, Berlin, 2007, pp. 1050–1057.
[32] M. Hunter, H. K. Kim, W. Suh, R. Fujimoto, J. Sirichoke, M. Palekar, Ad hoc distributed dynamic data-driven simulations of surface transportation systems, Simulation 85 (4) (2009) 243–255.
[33] B. Mitchell, L. Yilmaz, Symbiotic adaptive multisimulation: An autonomic simulation framework for real-time decision support under uncertainty, ACM Transactions on Modeling and Computer Simulation 19 (1) (2008) Article 2.
[34] Y. Huang, A. Verbraeck, A dynamic data-driven approach for rail transport system simulation, in: Proceedings of the 2009 Winter Simulation Conference, IEEE, Omnipress, 2009, pp. 2553–2562.
[35] B. Siegmund, Data Analysis: Statistical and Computational Methods for Scientists and Engineers, 3rd Edition, Springer-Verlag, New York, Berlin, 1998.
[36] C. Chatfield, The Analysis of Time Series: An Introduction, 6th Edition, CRC Press, 2003.
[37] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, From data mining to knowledge discovery in databases, American Association for Artificial Intelligence Magazine 17 (3) (1996) 37–53.