Research on the Dynamic Data-driven Application System Architecture for Flight Delay Prediction

Haiyan Chen *
College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics
Nanjing, P.R. China
Email: chenhaiyan@nuaa.edu.cn

Jiandong Wang
College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics
Nanjing, P.R. China
Email: aics@nuaa.edu.cn

Lirong Feng
Department of Mech. Eng., Colorado State Univ, Fort Collins,USA
Email: lrfeng@engr.colostate.edu

***Abstract*—Flight delay prediction remains an important research topic due to the dynamics of the flight operating process. To solve this problem, a dynamic data-driven approach from the control area has been introduced, in which real-time data was collected and injected into the prediction process to get more accurate and reliable results. In the case of predicting the landing delays of consecutive arrival flights, delay propagation was analyzed to establish the corresponding state space model. Then, dynamic data-driven prediction architecture for flight delay and the prediction steps of on this architecture were presented. Several experiments were carried out on historic flight data to validate the performance of this solution. Results show that the accuracy is high, and not sensitive to the number of the predicted flights. Therefore, the solution has good predictive stability and reliability.**

***Index Terms*—dynamic data-driven application system, flight delays prediction, parameter estimation, data assimilation, Kalman filters**

I. INTRODUCTION

As a result of excessive demand for air transportation, flight delay becomes an urgent problem that exacerbates national transportation capacity limitations. When a flight is delayed, it will probably affect the successive flight's on-schedule arrival or departure, and indirectly affect more downstream flights and airports. This is called delay propagation. However, if we can predict the state of flight in real-time, appropriate measures could be taken to reduce propagation effects, and also economic losses. Therefore, the real-time predictions of flight delay and propagation have great practical significance.

Over the past decade, studies have been focused on analyzing flight delay factors [1], delay propagation models [2], delay patterns [3-4], mitigating delays [5] and other issues. Deterministic models are commonly used in delay prediction. For example, one of the models is to estimate delays according to flight schedule. Models like this usually ignore random factors such as unexpected events and queuing. Prediction models based on random density functions of seasonal trends, daily propagation and daily delay [3-4] (that to a certain extent reflect the overall patterns of flight delays) are insufficient in capturing variations in individual flight delay, and can't be applied to predict the real-time state of each flight. Real-time prediction of flight delay is the state estimation process for a dynamic system in essential. The flight operation process is monitored in order to collect data in real time, which provides an opportunity to apply a dynamic data-driven application system (DDDAS) paradigm [6], which is the latest research area in system control, to real-time flight delay prediction.

In section II of this paper, the DDDAS paradigm and its application in transportation simulation were reviewed. Challenges of applying DDDAS were also discussed. Delay propagation among consecutive arrival flights was analyzed and a flight delay state space model was established in section III. In section IV, architecture of the dynamic data-driven system for flight delay prediction and the prediction process based on the Kalman filter were presented. Experiments were carried to validate the proposed architecture in section V.

II. DDDAS

Manuscript received January 1, 2011; revised June 1, 2011; accepted July 1, 2011.
*corresponding author

doi:10.4304/jsw.7.2.263-268

A. Introduction to DDDAS

In traditional system simulation, static data collected beforehand is input into a well-designed mathematical model to predict system states in the present or the future. In non-dynamic systems, it's feasible to use the traditional method of simulation. For dynamic systems, as the static input data can not capture the real-time changes in dynamic processes in a timely manner, the simulation results are often very different from the measured data, what leads to prediction failures. To compensate for the incompleteness of the system models, real-time information that can represent the actual state of the system can dynamically be added to the running simulations to provide more accurate and effective real-time predictions. For this purpose, DDDAS was proposed by the US's National Science Foundation (NSF) around 2000 [7] as a new paradigm for simulation and prediction applications.

DDDAS is a symbiotic feedback control system, which can dynamically steer the simulations based on the real-time measurement data, and in reverse, can dynamically employ simulations to control and guide the measurements, to determine when, where, and how it is best to gather additional data. This system promises more accurate analysis and prediction, more precise controls, and more reliable outcomes, which will improve modeling technologies, advance prediction capabilities of simulation systems, and enhance efficiency and effectiveness of measurement infrastructures.

DDDAS is an emerging and promising technology and has been applied to a variety of engineering and science practices in recent years [8], including crisis management, environmental science, disaster forecasting, engineering design and control, industrial manufacturing, medical, biotechnology, finance and trade. Transport simulation and air traffic management are further important application areas.

B. DDDAS in Transportation

Current application studies on DDDAS in the transportation area are mainly from NSF supported projects.

Reference [9] presented a hierarchical DDDAS architecture, including coupled in-vehicle, roadside, and traffic management center simulations, to apply dynamic data-driven simulation to monitor and manage surface transportation systems. However, the implementation and effectiveness evaluation of this architecture were not described.

Reference [10] looked into the use of a dynamic data-driven approach for surface transportation simulation to create an accurate estimate of the evolving state of transportation systems using real-time roadway data aggregated at various update intervals. Experiments based on a microscopic surface traffic simulation model show that simulation based on inflow data aggregated over a short time interval is capable of providing a superior representation of the real world over longer aggregate intervals. However, the perceived improvements are minimal under congested conditions and most pronounced under un-congested conditions.

Reference [11] discussed the potential benefits and requirements of dynamic data-driven simulation in rail systems, placing emphasis on automated model reconfiguration, calibration, and validation through the use of data analysis methods. A process model for data-driven calibration and validation was proposed, where the model output and the real measurement data were continually compared, and the model parameters would be updated if the deviation exceeded a predefined threshold. However, the proposed model wasn't implemented and case studies were not given in the paper.

Reference [12] reported on real data testing of a real-time freeway traffic state estimator, with a particular focus on its adaptive capabilities. The pursued general approach to the real-time adaptive estimation of the complete traffic state in freeway stretches or networks is based on stochastic macroscopic traffic flow modeling and an extended Kalman filter. Advantages are demonstrated via suitable real data testing. The achieved testing results are satisfactory and promising for subsequent applications.

Reference [13] proposed a dynamic data-driven multi-agent simulation framework for maritime traffic, with focus on the function description of each agent. However, there was no further discussion on how to collect and use the real-time data to estimate the runtime state of the vessel, and the proposed framework was too simple to be applied in the real world conditions.

Reference [14] presented an airline-flight-delay-predicting DDDAS – Flight Cast, which was the project of the DDDAS curriculum at Wyoming University. It was meant to accurately predict the probability of delay or cancellation of a flight. The function of the components in the framework, data used in the prediction and its collection scenario were presented in the paper. However, specific issues, such what kind of prediction model was to be used in certain applications, were not discussed and real data validation was not carried out to evaluate the performance of the proposed framework.

C. Challenges of Applying DDDAS

Existing studies on applying DDDAS in various types of transport provides useful guidelines to construct an integral DDDAS for flight delay prediction. From the studies reviewed above, challenges of applying the DDDAS paradigm can be summarized in the following issues:

(1) Computable prediction model. As the basis for DDDAS, a computable prediction model must first be established according to the system behavior, so as to describe the relationship between states, as well as the relationship between states and the measurable data, in a mathematical way.

(2) Data assimilation algorithm. According to the complexity of the prediction model, a certain data assimilation algorithm should be chosen to dynamically integrate the measurable data into the model and update the priori estimate to achieve more accurate prediction.

Focusing on these two problems, a case study on the landing delay prediction of consecutive arrival flights was carried out to demonstrate the key steps in applying DDDAS for flight delay prediction.

III STATE SPACE MODEL OF FLIGHT DELAY

The state space model is a mathematical model created by applying the state space analysis method in dynamic systems. There are two equations in a state space model: the process equation describes the evolution of the state variables of the dynamic system and the measurement equation represents how measurement data relate to the state variables. By using the state space model, immeasurable state variables can be incorporated into the measurable model to get an updated state estimate.

A. Delay Propagation of Consecutive Arrival Flights

As the runway cannot be used for more than one aircraft simultaneously, for any two consecutive arrival flights, delay effects will probably propagate to the succeeding flight if the previous flight is delayed. This is called the delay propagation chain. Here, we suppose that delay that occurs before the ready-for-landing signal is the cumulative delay, and the delay that occurs after the ready-for-landing signal is landing delay. The sum of these two delays is the arrival delay of the flight. Let *d*, *l*, *c* and *p* denote the arrival delay, landing delay, cumulative delay and delay propagation respectively, while *m* and *b* denote the minimum time interval and the buffer between consecutive flights; *m* must be included in *b* and be satisfied. The process of delay propagation occurrence is demonstrated in Fig. 1, where SAT means scheduled arrival time.

Fig.1 shows that no delay propagation will occur to the succeeding flight if d_1 is not greater than the difference between b_2 and m, otherwise, delay propagation, $p_2=d_1-b_2+m$, will occur on the succeeding flight to ensure the minimum time interval. The landing delay of the succeeding flight can be expressed as:

$$l_2 = p_2 + \varepsilon_2 = l_1 + c_1 - b_2 + m + \varepsilon_2 . \quad (1)$$

Where ε_2 denotes the delay caused by the uncertainties. It can be found that landing delay is linearly spread between consecutive flights.

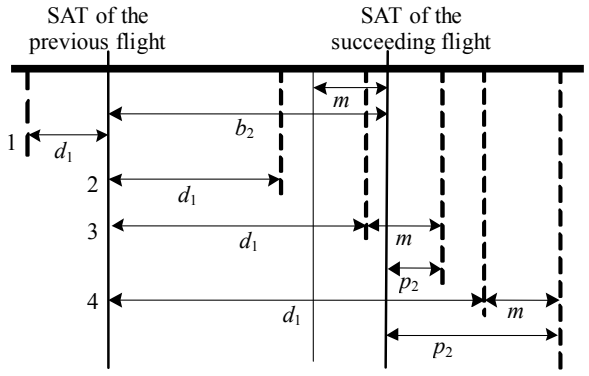


Figure 1. Delay propagation occurrence

B. Modeling of the State Space Model

Based on the analysis of landing delay, the state space model can be expressed as a piecewise linear model as follows:

$$x_k = \begin{cases} x_{k-1} + c_{k-1} - b_k + m + \varepsilon_k + w_k , & x_{k-1} + c_{k-1} \geq b_k - m \\ 0 + \varepsilon_k + w_k , & x_{k-1} + c_{k-1} < b_k - m \end{cases} \quad (2)$$

$$y_k = x_k + v_k \quad . \quad (3)$$

Where, (2) is the process equation, (3) is the measurement equation, x_k, and c_k, denote the landing delay and cumulative delay of the *k*th flight, b_k is the buffer between the *k*-1th and the *k*th flights, and y_k denotes the real-time measurement data. w_k and v_k denote the process and measurement noise, respectively; both are random white noise, that can be generated from the predefined variances. Since the relationships between the uncertainties and flight delays are not to be represented by any mathematical models, which leave the calculation of ε_k to be a problem in the establishment of the state-space model. Intelligent data mining algorithms can be used here to learn the model of the uncertainties effect from large amounts of historical data, so that ε_k under current conditions can be estimated by online update of the model using real-time data. ε_k should be defined by the expert in extreme circumstances, such as air controlled, snowstorm, etc.. The algorithm employed in the online estimate of ε_k and the validation of its generalization have been discussed in detail in reference [15], where the finite Gaussian Mixture model was applied to present the effect pattern of the uncertainties, a genetic EM algorithm was used to search to maximum likelihood estimation of the parameters, and the final model had generalization performance of over 90% in the validation.

IV DYNAMIC DATA-DRIVEN PREDICTION FOR FLIGHT DELAY

A. Dynamic Data-driven Application System Architecture for Flight Delay

Based on the state space model of landing delay, the dynamic data-driven application system architecture can be constructed, as shown in Fig. 2.

There are four main components in the architecture:

- The data acquisition and processing module collects and trims the real-time data, such as flight data, arrival schedule, airport conditions, weather conditions, etc., to for provide the online parameter estimation module and data assimilation module with available input data.
- The online parameter estimation module estimates the parameters c_k, b_k, w_k, v_k, ε_k in real-time conditions to improve the adaptability of the state space model.
- The system state space model, as the kernel of the architecture, calculates the landing delay priori estimate of the succeeding flight, according to the process equation and the results of on-line parameter

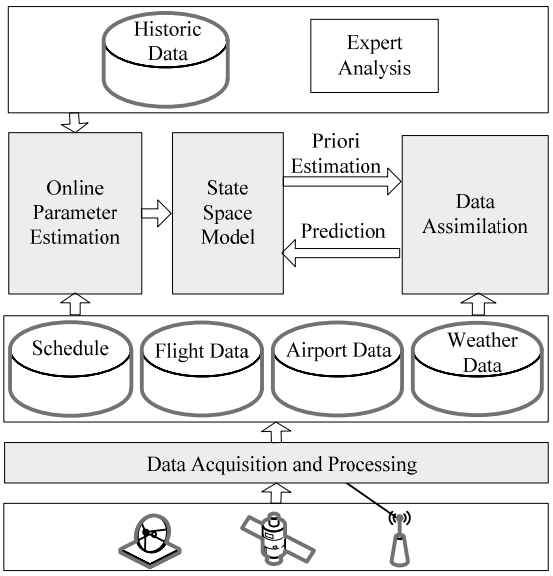


Figure 2. Dynamic Data-driven Application System Architecture for Flight Delay

estimation, and provides the input data to the data assimilation module.

- The data assimilation module updates the landing delay priori estimate with real-time measurement data. This is the essence of a dynamic data-driven approach. There are two methods available for data assimilation, the Kalman filter [16] for the state estimation of linear systems, and particle filter [17] for the state estimation of nonlinear non-Gaussian systems, as well as the extensions form of the two filters. Here we adopt the Kalman filter to complete data assimilation.

B. State Prediction Based on Kalman Filter

The Kalman filter is an optimal autoregressive data processing algorithm widely used in solving navigation, control, data integration and other problems. There are two parts to the Kalman filter: the time updating part which calculates the current state and error covariance in time to provide the priori estimate for next state, and the measurement-updating part which assimilates the priori estimate and the new arrival measurement data to produce an improved posteriori estimate, as a feedback to the model. To the linear system state space model $\begin{cases} x_k = Ax_{k-1} + Bu_{k-1} + w_k \\ y_k = Hx_k + v_k \end{cases}$, the process of state estimation and prediction based on the Kalman filter can be described as following steps:

Step1 Initial: giving state variable x_0 and its covariance P_0;

Step2 Time updating: calculating the priori estimates of x_k and P_k by expression (4) and (5) respectively.

$$x_{k|k-1} = Ax_{k-1} \tag{4}$$

$$P_{k|k-1} = P_{k-1} + Q_k \tag{5}$$

Where,

$$Q_k = \mathrm{var}(w_k)\,. \tag{6}$$

Step3 Measurement updating: updating the priori estimates of x_k and P_k by expression (7) when new measurement data y_k arrives, to get the posterior estimate of x_k and P_k.

$$x_k = x_{k|k-1} + K_k(y_k - Hx_{k|k-1}) \tag{7}$$

Where, K is the Kalman gain and can be calculated by the following expression.

$$K_k = \frac{P_{k|k-1}H'}{HP_{k|k-1}H'R_k} \tag{8}$$

Where,

$$R_k = \mathrm{var}(v_k)\,. \tag{9}$$

Step4 repeat Step2 and Step3 to carry forward the predictions.

C. Realization of Delay State Prediction

Based on the state space model defined in section III, realization of the state estimates and predictions were carried out on Matlab 7.1, as follows:

Define the variables: N, m, R, Q, $x(0)$, $P(0)$
For $k = 1{:}N$
(1) Calculate the priori estimate of the state
A=0;
$x(k)$=$\varepsilon(k)$;
If (x(k-1)+c(k-1)-b(k)+m >=0)
A=1;
$x(k)$= $x(k)$+ $x(k$-1)+c(k-1)-b(k)+m;
end
$P(k)$=A*P(k-1)*A'+Q;
$K(k)$=$P(k)$/($P(k)$+R);
(2) Update the priori estimate to get the posterior estimate
$x(k)$=$x(k)$+$K(k)$*($yc(k)$–$y(k)$);
$P(k)$=(1-$K(k)$)*$P(k)$;
$y(k)$= $x(k)$;
(3) Calculate the errors and root mean square errors
$e(k)$= $yc(k)$–$y(k)$;
$rmse(k)$=sqrt((norm(yc-y))^2/k);
end

V EXPERIMENTS AND ANALYSIS

To validate the performance of the proposed dynamic data-driven prediction for flight delay, a series of experiments were carried out for different considerations, such as prediction accuracy, noise effect, and steps of continuous predictions. Flight data, arrival schedules, observed landing times and other useful data were extracted from the flight operation records of a domestic hub airport to construct the data set for the experiments.

Experiment 1 was to verify the accuracy of the dynamic data-driven flight delay prediction and the impact on prediction accuracy when using different (R, Q). The parameters are set as follows: $x(0)$ = 8, $P(0)$ = 1, m= 4, (R, Q)=(1, 4). Fig. 3 demonstrates the landing delay predictions (the posteriori estimates) of 50

consecutive arrival flights, given by a particular run of

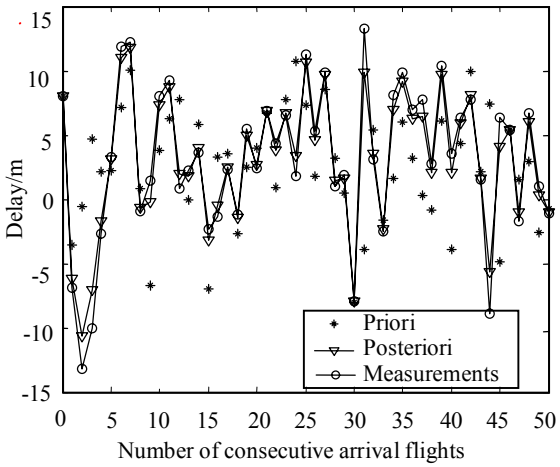


Figure 3. Landing Delay Predictions of 50 Consecutive Arrival Flights

the program. The final Root Mean Square Errors (RMSEs) of priori estimates and posteriori estimates when using five pairs of experiential (R,Q) are shown in Table 1.

RMSE comparisons show that prediction accuracy can be greatly improved by injecting the observed real-time data into the prediction process, which is verified by the prediction experiments. Both prove the effectiveness of the proposed dynamic data-driven prediction architecture for flight delay. It can also be seen that (R,Q) has a significant impact on the prediction accuracy. However, as errors between the observed value and the immeasurable true value always exist in practice, measurement accuracy should be improved to collect more reliable observed values, and (R,Q) should be estimated and adjusted real-time according to the characteristics of the predicted process, so as to obtain more accurate and reliable prediction.

Experiment 2 was to study the prediction accuracy at different numbers of consecutive arrival flights. Parameters are set as follows: $x(0) = 8$, $P(0) = 1$, $m= 4$, (R, Q)=(1, 4). The RMSEs of the delay predictions at 20, 30 and 50 consecutive flights are listed in Table 2.

It can be found from Table 2 that the number of consecutive flights has little effect on the prediction accuracy. That is, the proposed dynamic data-driven approach for flight delay prediction has good stability.

TABLE II. RMSES AT DIFFERENT (R,Q)

(R,Q)	RMSEs	
	Priori Estimates	Predictions
(1,1)	6.3555	2.6563
(1,4)	5.1724	1.0046
(1,9)	6.5458	1.2130
(0.01,9)	8.9549	0.0099
(4,9)	8.3408	2.2920

TABLE I. RMSES AT DIFFERENT NUMBERS OF CONSECUTIVE FLIGHTS

N	REMSs
20	0.649
30	0.794
50	1.0046

VI CONCLUSIONS AND FURTHER WORK

A dynamic data-driven approach for flight delay prediction has been presented in this paper to assimilate real-time measurement data with the priori estimates received from the system state space model, so as to refine the predictions dynamically. Case studies demonstrate that the proposed dynamic data-driven prediction approach has high prediction accuracy, and the prediction accuracy is almost insensitive to the number of consecutive flights. For further use of this dynamic data-driven prediction in air transportation, the flight state transition patterns at various stages should be analyzed with the existing measured data to establish computable state space models, based on which additional states of the flights can be estimated real-time. This will provide strong support to the airports or airlines to make reasonable decisions and arrangements to reduce flight delays

Further work can focus on the following three issues: how to improve the state space model to employ more real-time data, how to estimate the immeasurable variables using the existing or measurable data, and how to adjust the noise variances adaptively.

ACKNOWLEDGMENT

This research is funded by the High Technology Research and Development Program of China under Project 2006AA12A106. Authors would like to thank the anonymous domestic airline which provided historical flight information.

REFERENCES

[1] K. F. Abdelghany, S. S. Shah, S. Raina and A. F. Abdelghany, "A model for projecting flight delays during irregular operation conditions", *J. Air Transport Manag.,* vol.10, no.6, pp. 385-394, 2004. doi:10.1016/j.jairtraman.2004.06.008

[2] C. L. Hsu, C. C. Hsu and H. C. Li, "Flight Delay Propagation, Allowing for Behavioral Response", *Int. J. Crit. Infrastruct*, vol.3, no. 3/4, pp. 301-326, 2007. doi:10.1504/IJCIS.2007.014113

[3] Y. Tu, M, Ball, and W. Jank, "Estimating Flight Departure Delay Distributions—A Statistical Approach With Long-Term Trend and Short-Term Pattern", *J. Amer. Statistical Assoc.*, vol.103, no.481, pp. 112-125, 2008. doi: 10.1.1.132.1147

[4] M. Abdel-Aty, C. Lee, Y. Bai, X. Li and M. Michalak, "Detecting Periodic Patterns of Arrival Delay", *J Air*

Transport Manag., vol.13, no.6, pp. 355-361, 2008. doi:10.1016/j.jairtraman.2007.06.002

[5] S. AhmadBeygi, A. Cohn and M. Lapp, "Decreasing Airline Delay Propagation by Re-allocating Scheduled Slack", *IIE Trans.*, vol.42, no.7, pp. 478-489, 2010. doi: 10.1080/07408170903468605

[6] F. Darema, "Dynamic Data Driven Application Systems: A New Paradigm for Application Simulations and Measurement", in: *Proc. of the Intl Conf. on Computational Science*, Poland, 2004. doi:10.1.1.104.5449

[7] F. Darema, "Dynamic Data Driven Application Systems", NSF Workshop Report. http://www.dddas.org/nsf-workshop-2000/workshop_report.pdf, 2010-10-12

[8] F. Darema, "Introduction to the ICCS 2007 Workshop on Dynamic Data Driven Applications Systems", in: *Proc. of the Intl Conf. on Computational Science*, China, 2007. doi: 10.1007/978-3-540-72584-8_125

[9] R. M. Fujimoto, R. Guensler, M. Hunter, H. K. Kim, J. Lee, J. Leonard, et al, "Dynamic Data Driven Application Simulation of Surface Transportation Systems", in: *Proc. of the Intl Conf. on Computational Science*, UK, 2006. doi: 10.1007/11758532_57

[10] M. Hunter, R. M. Fujimoto, and W. Suh, "An Investigation of Real-time Dynamic Data Driven Transportation Simulation", in: *Proc. of the 2006 Winter Simulation Conference*, USA, pp. 1414-1421, 2006. doi: 10.1145/1218112.1218369

[11] Y. Huang, M. D Seck and A. Verbraeck, "Towards Automated Model Calibration and Validation in Rail Transit Simulation", *Procedia Comput. Sci*., vol.1 no.1 pp. 1259-1265, 2010. doi: 10.1016/j.procs.2010.04.140

[12] Y Wang, M Papageorgiou, A Messmer, P. Coppola, A. Tzimitsi, A. Nuzzolo, "An Adaptive Freeway Traffic State Estimator", *Automatica*, vol.45, no.1, pp. 10-24, 2009. doi: 10.1016/j.automatica.2008.05.019

[13] Y J Xiao, H Zhang and S Li, "Dynamic Data Driven Multi-Agent Simulation in Maritime traffic", in: *Proc. of the Inlt. Conf. on Computer and Automation Engineering*, Thailand, pp. 234-237, 2009. doi: 10.1109/ICCAE.2009.17

[14] R Hyatt, D Bansal and S Chakraborty, "Flight Cast – An Airline Flight Delay Prediction DDDAS", in: *Proc. of the Inlt. Symp.* on Distributed Computing *and Application to Business, Engineering and Science*, China, pp. 85-88, 2007. http://dcabes.meeting.whut.edu.cn/.../DCABES%202007%20Proceedings%20Volume%20I.pdf, 2011-1-5

[15] H Y Chen, J D Wang and T Xu, "Modeling of Flight Delay State-space Model Based on Genetic EM Algorithm", *Transactions of NUAA*, China, in press, 2011.

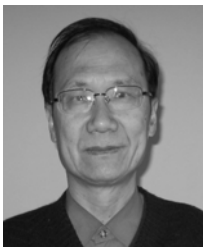
[16] R E Kalman, R S Bucy, "New Results in Linear Filtering and Prediction Problems", *J. Basic Eng.*, vol.83, no.1, pp. 95-108, 1961. doi: 10.1.1.129.6247

[17] E Bølviken, P J Acklam and N Christophersen, "Monte Carlo Filters for Non-linear State Estimation", *Automatica*, *UK*, vol.37, no.2, pp. 177-183, 2001. doi: 10.1016/S0005-1098(00)00151-5

Haiyan Chen was born in Changzhou, P.R. China, in 1979. She received the B.S. degree in Computer Science and Technology from the Nanjing University of Aeronautics and Astronautics (NUAA), Nanjing, P.R. China, in 2002, and the M.S. degree in Computer Application from NUAA, Nanjing, P.R. China, in 2005.

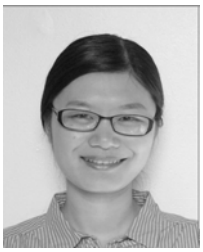
From 2005 to 2006, she worked as a Teaching Assistant at College of Computer Science and Technology at NUAA and as a Lecture since 2007. From 2006 to 2010, she worked as a Data Analyst for the project of High Technology Research and Development Program of China No. 2006AA12A106.

Haiyan Chen is currently pursuing her Ph.D. in system modeling and simulation research at college of computer science and technology of NUAA. Her study concerns data mining, modeling and simulation.



Jiandong Wang was born in Shuyang, P. R. China, in 1945. He received the M.S. degree in Electrical Engineering from Shanghai Jiao Tong University. He was a visiting scholar at the University of Ottawa, Canada, from 1990 to 1991.

He is currently Professor and Doctoral Students Adviser in the College of Computer Science and Technology at NUAA. His main research interests include artificial intelligence, data mining and information security.



Lirong Feng was born in Nantong, P.R. China, in 1979. She received the B.S. degree in Computer Science and Technology from the Nanjing University of Aeronautics and Astronautics, Nanjing, P.R. China, in 2002, and the M.S. degree in electrical engineering from the Colorado state university, Fort Collins, Colorado, in 2007.

From 2007 to 2009, she worked with Sallie Mae, Inc., the leading student loan corporation, as a financial risk analyst, focused on the North America market. Prior to joining Sallie Mae, she worked as a research assistant at Colorado State University, focused on weather radar signal processing in 2005, and as a student operation scientist for the NASA CloudSat project in 2006.

Lirong Feng is currently pursuing her Ph.D. in industrial engineering and operations research at the m echanical engineering department, Colorado state university. Her research interest is on the modeling and optimization in a variety of areas includes experimental design, risk theory, and inventory management.