

**DATA SCIENCE APPLICATION IN INTELLIGENT
TRANSPORTATION SYSTEMS: AN INTEGRATIVE
APPROACH FOR BORDER DELAY PREDICTION AND
TRAFFIC ACCIDENT ANALYSIS**

by

Lei Lin
January 5, 2015

A dissertation submitted to the
Faculty of the Graduate School of
the University at Buffalo, the State University of New York
in partial fulfillment of the requirements for the
degree of

Doctor of Philosophy

Department of Civil, Structural, and Environmental Engineering

UMI Number: 3683052

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



UMI 3683052

Published by ProQuest LLC (2015). Copyright in the Dissertation held by the Author.

Microform Edition © ProQuest LLC.

All rights reserved. This work is protected against unauthorized copying under Title 17, United States Code



ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

Copyright by

Lei Lin

2015

All rights reserved.

PREVIEW

ACKNOWLEDGEMENTS

First and foremost I would like to offer my greatest gratitude to my two supervisors, Dr. Adel W. Sadek and Dr. Qian Wang, who have supported me throughout my research studies with their patience and knowledge. Both of them have given me continuous encouragement and guidance on my research works in the past four and a half years. They always help me with useful suggestions to make research plans; they always respond very fast to my emails about any difficulties I met in the studies no matter how busy they are; and they always revise my paper sentence by sentence even a small typo in a figure. Without their help, this dissertation would not have been possible. I feel so fortunate to be one of their Ph.D. students. The moments that we have thought-provoking discussions, get promising experiment results, and work hard together on one paper will always be remembered by me.

Second, my special thanks to my family. My parents Jianzhou Lin and Jianrong Chu, who unconditionally support every decision I have made in my life, are my source of energy. My wife Yan Li, she always takes good care of me and also gives me huge help in getting my minor degree, the Master's Degree of Computer Science.

Finally, I would also like to thank my committee members, Dr. Panagiotis Ch. Anastasopoulos and Dr. Qing He for accepting the invitation to serve on my committee. I also feel very grateful for their precious advice and great ideas towards my dissertation.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS	iii
TABLE OF CONTENTS	iv
LIST OF TABLES	xv
LIST OF FIGURES	xviii
ABBREVIATIONS	xxi
ABSTRACT	xxiii
CHAPTER 1 INTRODUCTION.....	1
1.1 Data Science in Intelligent Transportation Systems	1
1.2 ITS Data Introduction and Application.....	3
1.2.1 Fixed Sensors	3
1.2.2 Probe Vehicles	4
1.2.3 Connected Vehicles	4
1.2.4 Inspection and Toll Stations.....	5
1.2.5 Traffic Accidents	5
1.2.6 Social Media and Crowd Sourcing	7
1.2.7 Web-based Mapping Services.....	7
1.2.8 Computer Simulation Data	8
1.2.9 Weather Data	9
1.3 ITS Data Analysis	9
1.3.1 The ITS Data Analysis Process.....	10

1.3.1.1	Data Preprocessing.....	12
1.3.1.2	Data Fusion	13
1.3.1.3	Data “Width and Depth” Reduction	14
1.3.1.4	Model Development.....	16
1.3.1.5	Model Application	18
1.3.2	ITS Data Analysis Models	18
1.3.2.1	Statistical Models.....	19
1.3.2.2	Machine Learning and Data Mining.....	20
1.3.2.3	Other Types of Models	21
1.3.2.4	Summary	22
1.4	Research Overview	23
1.4.1	Two Case Studies.....	23
1.4.1.1	Niagara Frontier International Border Crossing Delay Prediction.....	23
1.4.1.2	Traffic Accident Data Analysis	26
1.4.2	Five Research Subtopics	27
1.4.2.1	Short-term Traffic Volume Prediction Model	27
1.4.2.2	Queueing Model.....	29
1.4.2.3	Traffic Accident Hotspots and Clearance Time Analysis	30
1.4.2.4	Real-time Traffic Accident Risk Prediction	31
1.4.2.5	Traffic Accident Duration Prediction	31

1.5	Research Contributions.....	32
1.5.1	Dissertation Contributions to the ITS Data Analysis Process	34
1.5.1.1	Integration of Data “Width” Reduction Methods and Model Development.....	34
1.5.1.2	Integration of Data “Depth” Reduction Methods and Model Development.....	35
1.5.1.3	Integration of Data Diagnosis and Short-term Traffic Volume Model Development.....	35
1.5.1.4	Model Combination of SARIMA and SVR for Short-term Traffic Volume Prediction.....	36
1.5.1.5	Model Application - an Android Smartphone Application for Niagara Frontier Border Crossing	37
1.5.2	The Dissertation Contributions to ITS Data Models	37
1.5.2.1	Integration of Dynamic Time Warping (DTW) and Spinning Network (SPN) for Short-term Traffic Volume Prediction.....	38
1.5.2.2	Integration of Data-driven Models and Analytical Methods for Border Delay Prediction.....	38
1.5.2.3	Integration of M5P tree and HBDM for Traffic Accident Duration Prediction.....	39
1.5.3	Novelty Discussion	39
1.6	Dissertation Organization	40

CHAPTER 2 LITERATURE REVIEW	43
2.1 Border Crossing Delay Prediction	43
2.1.1 Short-term Traffic Volume Prediction.....	43
2.1.1.1 Data “Depth” Decreasing.....	44
2.1.1.2 Data Diagnosis	45
2.1.1.3 Model Development.....	46
2.1.2 Queueing Model.....	49
2.1.3 Other Border Crossing Delay Studies.....	52
2.2 Traffic Accident Data Analysis	53
2.2.1 Data “Width” Decreasing	53
2.2.2 Data “Depth” Decreasing.....	54
2.2.3 Model Development.....	56
2.2.3.1 Hotspots Analysis	56
2.2.3.2 Real-time Traffic Accident Risk Prediction	58
2.2.3.3 Traffic Accident Duration Prediction	60
CHAPTER 3 SHORT-TERM TRAFFIC VOLUME PREDICTION.....	64
3.1 Data “Depth” Decreasing and Model Combination for Border Crossing Traffic Prediction.....	64
3.1.1 Methodology	64
3.1.1.1 Dataset Grouping	64
3.1.1.2 Combination of SARIMA and SVR	65
3.1.2 Modeling Dataset.....	70

3.1.3 Modeling Development and Results.....	73
3.1.3.1 Prediction Accuracy Measure.....	73
3.1.3.2 SARIMA Model.....	73
3.1.3.3 SVR Model.....	76
3.1.3.4 Multi-model Combined Forecasting Method.....	78
3.2 On-line Prediction of Border Crossing Traffic Using DTW-SPN.....	81
3.2.1 Methodology.....	81
3.2.1.1 Spinning Network method (SPN).....	82
3.2.2 Datasets.....	84
3.2.3 Model Development.....	85
3.2.3.1 Enhanced SPN.....	85
3.2.3.2 SPN parameters.....	87
3.2.3.3 SARIMA model and SVR model.....	90
3.2.4 Evaluation Results.....	90
3.2.4.1 Comparisons of the four models based on the classified dataset.....	91
3.2.4.2 Comparison of the four models based on the non-classified dataset.....	95
3.2.4.3 Impact of data classification.....	98
3.2.4.4 Running time comparison of the four models.....	100
3.3 Evaluating Short-term Traffic Prediction Models Based on Multiple Datasets and Data Diagnosis Measures.....	102

3.3.1 Methodology	102
3.3.1.1 Data Diagnosis	102
3.3.1.2 Short-term Traffic Volume Prediction Models.....	105
3.3.2 Modeling Datasets	106
3.3.3 Data Diagnosis Results	107
3.3.3.1 Delay Time and Embedding Dimension.....	107
3.3.3.2 Approximate Entropy.....	108
3.3.3.3 Time Reversibility of Surrogate Data.....	109
3.3.3.4 Hurst Exponent	110
3.3.4 Model Development.....	111
3.3.4.1 SARIMA Models.....	111
3.3.4.2 K-NN Models.....	113
3.3.4.3 SVR Models.....	115
3.3.4.4 Comparison of Model Performance.....	116
3.4 Conclusions.....	118

CHAPTER 4 SOLUTIONS OF TRANSIENT MULTI-SERVER

QUEUEING MODELS 123

4.1 The Peace Bridge Case Study	123
4.1.1 Estimation of Arrival and Service Process Distributions - Model 1	
.....	124
4.1.2 Estimation of Arrival and Service Process Distributions - Model 2	
.....	125

4.1.2.1	Batch Markov Arrival Process.....	126
4.1.2.2	Phase Type Distribution for Service Process.....	128
4.2	Methodology.....	130
4.2.1	<i>M/Ek/n</i> queueing models.....	131
4.2.1.1	System state description.....	132
4.2.1.2	State transition probabilities	134
4.2.1.3	State-to-State Transitions and Transient Solution Calculations.....	141
4.2.1.4	Performance Measurement Calculations	145
4.2.2	BMAP/PH/n queueing model	146
4.2.2.1	System state description.....	147
4.2.2.2	State transition probabilities	148
4.2.2.3	State Probabilities and Transient Solution Calculations.	152
4.2.3	The Baseline micro-simulation VISSIM model	155
4.3	Results.....	159
4.3.1	Validation Results.....	159
4.3.2	Sensitivity Analysis	162
4.3.2.1	Impact of an increasing travel demand, λ	162
4.3.2.2	Impact of opening additional service stations, n	164
4.3.2.3	Impact of changes in the mean service time μ	165
4.3.3	Optimal Operating Policies.....	166

4.3.3.1	Optimization Problem Formulation	166
4.3.3.2	Optimization Problem Results	168
4.4	Conclusions.....	171
 CHAPTER 5 ANDROID SMARTPHONE APP FOR TORONTO		
BUFFALO BORDER WAITING.....174		
5.1	Methodology behind the App	176
5.1.1	Integration of Data-driven Model and Analytical Model for Future Border Crossing Waiting Time	176
5.1.1.1	Border Crossing Traffic Volume Prediction Module	177
5.1.1.2	Transient Multi-server Queueing Module	177
5.1.2	Data-level Fusion for Current Border Crossing Waiting Time	178
5.2	Datasets	179
5.3	Innovative Features	179
5.3.1	Sharing Current Waiting Time Function	179
5.3.2	Utilizing Historical Waiting Time Function	181
5.3.3	Predicting Future Waiting Time Function	182
5.4	Comparison with Other Border Crossing Apps	186
5.5	Risks and Challenges	188
5.5.1	The Need for More Data	188
5.5.2	Crowd Sourcing	189
5.5.3	GPS Location	189
5.6	Conclusions.....	190

CHAPTER 6 DATA MINING AND COMPLEX NETWORKS

ALGORITHMS FOR TRAFFIC ACCIDENT ANALYSIS.....191

6.1 Methodology.....	191
6.1.1 Dataset Clustering.....	191
6.1.1.1 Modularity-based Community Detection	192
6.1.2 Hotspots and Clearance Time Analysis	195
6.1.2.1 Association Rule Learning.....	195
6.2 Dataset Processing.....	196
6.3 Results.....	199
6.3.1 Community Detection.....	199
6.3.2 Association Rule Analysis to Identify Hotspots	203
6.3.3 Association Rule Analysis to Identify Factors Affecting Incident Clearance Time.....	208
6.4 Conclusions.....	213

CHAPTER 7 NOVEL VARIABLE SELECTION FOR REAL-TIME

TRAFFIC ACCIDENT RISK PREDICTION214

7.1 Methodology.....	214
7.1.1 Variable Selection.....	214
7.1.1.1 Frequent pattern Tree (FP tree).....	215
7.1.1.2 Random forest.....	222
7.1.2 Models.....	224
7.1.2.1 k nearest neighbor (k-NN)	224

7.1.2.2	Bayesian Network.....	224
7.2	Modeling dataset.....	226
7.3	Model development and results	230
7.3.1	Variable importance calculation	230
7.3.2	k-NN	237
7.3.3	Bayesian network.....	239
7.4	Conclusions.....	242
CHAPTER 8 TRAFFIC ACCIDENT DURATION PREDICTION BASED		
ON M5P TREE AND HBDM		244
8.1	Methodology.....	244
8.1.1	M5P Tree	246
8.1.2	Hazard-based Duration Model (HBDM)	250
8.1.3	M5P-HBDM Model.....	252
8.2	Modeling Datasets	256
8.2.1	Virginia Traffic Accident Dataset.....	256
8.2.2	Buffalo Traffic Accident Detector	258
8.2.3	Accident Duration Characteristics	261
8.3	Model Development and Comparison	262
8.3.1	M5P Tree	262
8.3.2	Hazard-based Duration Model.....	267
8.3.3	M5P-HBDM Model.....	270
8.3.4	Significant Variables Comparison	275

8.3.5 Duration Prediction Comparison	278
8.4 Conclusions.....	283
CHAPTER 9 DISSERTATION CONTRIBUTION AND FUTURE	
RESEARCH	286
REFERENCE.....	297

PREVIEW

LIST OF TABLES

Table 1-1 Research Contributions Summary	33
Table 3-1 Prediction Performance of the SARIMA Model	75
Table 3-2 Prediction Performance of the SVR Model	78
Table 3-3 Models' Prediction Performance Comparison	80
Table 3-4 Model Performance for the Abrupt Points in the Classified Datasets	93
Table 3-5 Prediction Performance of SPN, SARIMA, and SVR for the Unclassified Data	96
Table 3-6 Computational Time Complexity of SPN, SARIMA, and SVR for the Unclassified Data	102
Table 3-7 Mutual Information Values With Respect to Time Delay	107
Table 3-8 Performance of SARIMA With Respect to Training Data Size	112
Table 3-9 Performance of SVR With Respect to Training Data Size	115
Table 3-10 Comparisons of Three Prediction Models for Four Datasets	117
Table 4-1 Notation for Queueing Models	134
Table 4-2 Results of Analytical Approach and Simulation Approach	160
Table 4-3 Impact of Increasing the Traffic Demand Level	162
Table 5-1 Prediction Performance of the Two-step Delay Prediction Model	183
Table 5-2 Comparison of TBBW with Other Border Crossing Apps	186
Table 6-1 Traffic Accident Variables in the I-190 Data	198
Table 6-2 Network Clusters With Respect to the Similarity Threshold	199

Table 6-3 Causative Factors and Their Probabilities in Each Cluster (%)	201
Table 6-4 Traffic Accident Types	203
Table 6-5 Rules on Hotspots from the Whole Dataset and the Clusters.....	204
Table 6-6 Rules on Clearance Time from the Whole Dataset and the Clusters	209
Table 7-1 Clustering Results for 5-minute and 10-minute Accident Training Datasets	231
Table 7-2 Supports of Items in 5-minute and 10-minute Accident Training Datasets	232
Table 7-3 Variable Importance Calculations Results based on FP Tree and Random Forest Methods.....	235
Table 7-4 Performance of k-NN for Different Variable Selection	237
Table 7-5 Prediction Performance of Bayesian Network with Different Variable Selection Strategies.....	240
Table 8-1 the Pseudo-Code of M5P-HBDM Algorithm and Comparison with M5P Tree	252
Table 8-2 Traffic Accident Variables in I-64 Dataset	256
Table 8-3 Traffic Accident Variables in I-190 Dataset	258
Table 8-4 Statistical Analysis of Accident Duration for I-64 and I-190.....	261
Table 8-5 AIC Values of HBDMs for I-64 and I-190 Training Datasets	268
Table 8-6 Log-normal AFT Models on I-64 Training Dataset.....	269
Table 8-7 Log-normal AFT Models on I-190 Training Dataset.....	270
Table 8-8 Log-normal AFT Models in M5P-HBDM of I-64 Training Dataset	271

Table 8-9 Log-normal AFT Models in M5P-HBDM of I-190 Training Dataset	273
Table 8-10 Significant Variables in M5P, HBDM and M5P-HBDM of I-64 Training Dataset.....	276
Table 8-11 Significant Variables in M5P, HBDM and M5P-HBDM of I-190 Training Dataset.....	277
Table 8-12 MAPEs of M5P Tree, HBDM Model and M5P-HBDM Model for I-64 and I-190 Testing Datasets	280
Table 8-13 Performances of M5P, HBDM and M5P-HBDM for Different Actual Duration Intervals of I-64 and I-190 Testing Datasets	281
Table 8-14 Percentage of Predictions Having a Difference within a Certain Tolerance and Mean Absolute Differences of M5P Tree, HBDM Model and M5P-HBDM Model for I-64 and I-190 Testing Datasets	282

PREVIEW

LIST OF FIGURES

Figure 1-1 ITS Data Analysis Process	11
Figure 1-2 Three Bridges at Niagara Frontier International Border	24
Figure 1-3 Framework of the Two-step Border Crossing Delay Prediction.....	26
Figure 3-1 Hourly Traffic Volume at the Peace Bridge on Different Days.....	71
Figure 3-2 Training Dataset Length and Model Performance	75
Figure 3-3 Traffic Volume Predictions of SARIMA, SVR, and the Combined Forecasting Method with Variable Weight.....	81
Figure 3-4 Spinning Network (SPN) (Revised based on Huang and Sadek, 2009)	83
Figure 3-5 Similarity between Time-dependent Sequences	86
Figure 3-6 Prediction Performance of the Four Models for Different Data Classes ...	91
Figure 3-7 Estimation Performances of Four Models for the Friday Group	94
Figure 3-8 Predictions of Four Models versus Actual Volumes in the Friday Group.	95
Figure 3-9 Estimation Performances of Four Models for the Whole Dataset	97
Figure 3-10 Hourly Volume Predictions versus Observations in the Whole Dataset..	97
Figure 3-11 Comparison of Model Performance between the Classified Case and the Unclassified Case.....	99
Figure 3-12 Percent of False Nearest Neighbors With Respect to Embedding Dimension.....	108
Figure 3-13 Test of Nonlinearity through Surrogate Data.....	110

Figure 3-14 Performance of k-NN With Respect to Input Data Vector Length (B) and the Number of Nearest Neighbors (k)	114
Figure 3-15 Performance of SVR With Respect to Input Data Vector Length D	116
Figure 4-1 Inter-arrival Time Distribution.....	125
Figure 4-2 Service Time Distribution.....	125
Figure 4-3 PH Distribution of the Border Service Process.....	130
Figure 4-4 an Example for Stages, Patterns and States	133
Figure 4-5 State to State Transition Process for the $M/E_{k=2}/3$ Queueing Model..	140
Figure 4-6 Queueing Model in VISSIM	157
Figure 4-7 Setting of Dwell Time Distribution in Stop Sign to Simulate Service Time	158
Figure 4-8 Delay Curve of 20 Minutes for $\lambda = 500 \text{ veh/h}$ and $\mu = 30\text{s}$ and $n = 3$	164
Figure 4-9 Delay at the End of 20 Minutes for Different Service Station Number n with $\lambda = 500 \text{ veh/h}$ and $\mu = 30\text{s}$	165
Figure 4-10 Service Station Number for Different Average Service Time of Erlang Distribution	166
Figure 4-11 Traffic Volume of every 20 minutes' Interval for a Whole Day	168
Figure 4-12 Total Cost of Optimizing the Queueing System for a Whole Day	169
Figure 4-13 Number of Open Booth of Optimizing the Queueing System for a Whole Day.....	170

Figure 5-1 Comparison of TBBW with the Other Ways to Share Border Waiting Time	176
Figure 5-2 Three Ways to Share Current Waiting Time.....	180
Figure 5-3 Three Ways to Utilize Historical Waiting Time	181
Figure 5-4 Predicted Border Crossing Waiting Time	183
Figure 5-5 Prediction Performance for the Peak Hours of 18:00-20:00 on April 22, 2014.....	185
Figure 6-1 Resulting traffic accidents network and community detection (<i>eth</i> = 7)	200
Figure 7-1 Frequent pattern (FP) tree.	217
Figure 7-2 Part of I-64 in Norfolk, Virginia.....	227
Figure 7-3 A part of the FP Tree for the 10-minute training dataset	234
Figure 7-4 Bayesian Network for 10 min Dataset Using Variables based on FP Tree	240
Figure 8-1 an Example of Tree Pruning Step	255
Figure 8-2 Density Distributions of Accident Duration for I-64 and I-190.....	261
Figure 8-3 M5P Tree Model for I-64 Training Dataset	263
Figure 8-4 M5P Tree Model for I-190 Training Dataset	265
Figure 8-5 M5P-HBDM Model for I-64 Training Dataset	271
Figure 8-6 M5P-HBDM Model for I-190 Training Dataset	273

ABBREVIATIONS

ApEn	Approximate Entropy
AVI	Automatic Vehicle Identification
AVL	Automatic Vehicle Location
BMAP	Batch Markovian Arrival Process
CI	Conditional Independence
DTW	Dynamic Time Warping
ELC	Equally Likely Combination
ELV	Equally Likely Vehicles
EM	Expectation and Maximization
ETC	Electronic Toll Collection
FCM	Fuzzy C-means Clustering Method
FHWA	Federal Highway Administration
FP tree	Frequent Pattern Tree
k-NN	k Nearest Neighbor
HBDM	Hazard-based Duration Model
IAAFT	Iteratively Amplitude Adjusted Fourier Transform
ILD	Inductive Loop Detector
ITS	Intelligent Transportation System
LCC	Latent Class Clustering
LGP	Linear Gaussian Process

LRD	Long Range Dependence
MAPE	Mean Absolute Percentage Error
$M/E_K/n$	Queueing Model with Exponential inter-arrival times and Erlang service times
NITTEC	Niagara International Transportation Technology Coalition
OPR	Object Purity Ratio
PH	Phase Types
ROPR	Relative Object Purity Ratio
SARIMA	Seasonal Autoregressive Integrated Moving Average Model
SPN	Spinning Network Method
SPSS	Statistical Package for Social Sciences
SVR	Support Vector Regression

ABSTRACT

With the great progress in information and communications technologies in the past few decades, intelligent transportation systems (ITS) have accumulated vast amounts of data regarding the movement of people and goods from one location to another. Besides the traditional fixed sensors and GPS devices, new emerging data sources and approaches such as social media and crowdsourcing can be used to extract travel-related data, especially given the wide popularity of mobile devices such as smartphones and tablets, along with their associated apps. To take advantage of all these data and to address the associated challenges, big data techniques, and a new emerging field called data science, are currently receiving more and more attention. Data science employs techniques and theories from many fields such as statistics, machine learning, data mining, analytical models and computer programming to solve the data analysis task. It is therefore timely and important to explore how data science may be best employed for transportation data analysis. In this doctoral study, an integrative approach is proposed for data science applications in ITS. The proposed approach constitutes to an integration of multiple steps in the data analysis process, or integration of different models to build a more powerful one. The integrative approach is applied and tested on two case studies: border crossing delay prediction and traffic accident data analysis.

For the first case study, a two-step border crossing delay prediction model is proposed, consisting of a short-term traffic volume prediction model and a