University Transportation Research Center - Region 2

Final Report

Techniques for Efficient Detection of Rapid Weather Change and Analysis of their Impacts on a Highway Network

Performing Organization: State University of New York (SUNY)

October 2017

Sponsor:
University Transportation Research Center - Region 2
Federal Highway Administration/U.S. Department of Transportation

University Transportation Research Center - Region 2

The Region 2 University Transportation Research Center (UTRC) is one of ten original University Transportation Centers established in 1987 by the U.S. Congress. These Centers were established with the recognition that transportation plays a key role in the nation's economy and the quality of life of its citizens. University faculty members provide a critical link in resolving our national and regional transportation problems while training the professionals who address our transportation systems and their customers on a daily basis.

The UTRC was established in order to support research, education and the transfer of technology in the field of transportation. The theme of the Center is "Planning and Managing Regional Transportation Systems in a Changing World." Presently, under the direction of Dr. Camille Kamga, the UTRC represents USDOT Region II, including New York, New Jersey, Puerto Rico and the U.S. Virgin Islands. Functioning as a consortium of twelve major Universities throughout the region, UTRC is located at the CUNY Institute for Transportation Systems at The City College of New York, the lead institution of the consortium. The Center, through its consortium, an Agency-Industry Council and its Director and Staff, supports research, education, and technology transfer under its theme. UTRC's three main goals are:

Research

The research program objectives are (1) to develop a theme based transportation research program that is responsive to the needs of regional transportation organizations and stakeholders, and (2) to conduct that program in cooperation with the partners. The program includes both studies that are identified with research partners of projects targeted to the theme, and targeted, short-term projects. The program develops competitive proposals, which are evaluated to insure the mostresponsive UTRC team conducts the work. The research program is responsive to the UTRC theme: "Planning and Managing Regional Transportation Systems in a Changing World." The complex transportation system of transit and infrastructure, and the rapidly changing environment impacts the nation's largest city and metropolitan area. The New York/New Jersey Metropolitan has over 19 million people, 600,000 businesses and 9 million workers. The Region's intermodal and multimodal systems must serve all customers and stakeholders within the region and globally.Under the current grant, the new research projects and the ongoing research projects concentrate the program efforts on the categories of Transportation Systems Performance and Information Infrastructure to provide needed services to the New Jersey Department of Transportation, New York City Department of Transportation, New York Metropolitan Transportation Council , New York State Department of Transportation, and the New York State Energy and Research Development Authorityand others, all while enhancing the center's theme.

Education and Workforce Development

The modern professional must combine the technical skills of engineering and planning with knowledge of economics, environmental science, management, finance, and law as well as negotiation skills, psychology and sociology. And, she/he must be computer literate, wired to the web, and knowledgeable about advances in information technology. UTRC's education and training efforts provide a multidisciplinary program of course work and experiential learning to train students and provide advanced training or retraining of practitioners to plan and manage regional transportation systems. UTRC must meet the need to educate the undergraduate and graduate student with a foundation of transportation fundamentals that allows for solving complex problems in a world much more dynamic than even a decade ago. Simultaneously, the demand for continuing education is growing – either because of professional license requirements or because the workplace demands it – and provides the opportunity to combine State of Practice education with tailored ways of delivering content.

Technology Transfer

UTRC's Technology Transfer Program goes beyond what might be considered "traditional" technology transfer activities. Its main objectives are (1) to increase the awareness and level of information concerning transportation issues facing Region 2; (2) to improve the knowledge base and approach to problem solving of the region's transportation workforce, from those operating the systems to those at the most senior level of managing the system; and by doing so, to improve the overall professional capability of the transportation workforce; (3) to stimulate discussion and debate concerning the integration of new technologies into our culture, our work and our transportation systems; (4) to provide the more traditional but extremely important job of disseminating research and project reports, studies, analysis and use of tools to the education, research and practicing community both nationally and internationally; and (5) to provide unbiased information and testimony to decision-makers concerning regional transportation issues consistent with the UTRC theme.

Project No(s):
UTRC/RF Grant No: 49198-39-28

Project Date: October 2017

Project Title: Techniques for Efficient Detection of Rapid Weather Change and Analysis of their Impacts on a Highway Network

Project's Website:
http://www.utrc2.org/research/projects/techniques-efficient-detection-rapid-weather-change

Principal Investigator(s):
Catherine T. Lawson
Associate Professor
Department of Geography and Planning
SUNY Albany
Albany, NY 12222
Tel: (518) 442-4775
Email: lawsonc@albany.edu

Feng Chen
Assistant Professor
Department of Computer Science
SUNY Albany
Albany, NY 12222
Tel: (518) 437-4940
Email: fchen5@albany.edu

Co Author(s):
Adil Alim
SUNY Albany
Albany, NY 12222

Joshi Aparna
SUNY Albany
Albany, NY 12222

Performing Organization(s):
State University of New York (SUNY)

Sponsor(s):
University Transportation Research Center (UTRC)
U.S. Department of Transportation/Federal Highway Administration

To request a hard copy of our final reports, please send us an email at utrc@utrc2.org

Mailing Address:

University Transportation Reserch Center
The City College of New York
Marshak Hall, Suite 910
160 Convent Avenue
New York, NY 10031
Tel: 212-650-8051
Fax: 212-650-8374
Web: www.utrc2.org

Board of Directors

The UTRC Board of Directors consists of one or two members from each Consortium school (each school receives two votes regardless of the number of representatives on the board). The Center Director is an ex-officio member of the Board and The Center management team serves as staff to the Board.

City University of New York

Dr. Robert E. Paaswell - Director Emeritus of UTRC
Dr. Hongmian Gong - Geography/Hunter College

Clarkson University
Dr. Kerop D. Janoyan - Civil Engineering

Columbia University
Dr. Raimondo Betti - Civil Engineering
Dr. Elliott Sclar - Urban and Regional Planning

Cornell University
Dr. Huaizhu (Oliver) Gao - Civil Engineering

Hofstra University
Dr. Jean-Paul Rodrigue - Global Studies and Geography

Manhattan College
Dr. Anirban De - Civil & Environmental Engineering
Dr. Matthew Volovski - Civil & Environmental Engineering

New Jersey Institute of Technology
Dr. Steven I-Jy Chien - Civil Engineering
Dr. Joyoung Lee - Civil & Environmental Engineering

New York Institute of Technology
Dr. Marta Panero - Director, Strategic Partnerships
Nada Marie Anid - Professor & Dean of the School of Engineering & Computing Sciences

New York University
Dr. Mitchell L. Moss - Urban Policy and Planning
Dr. Rae Zimmerman - Planning and Public Administration
Dr. Kaan Ozbay - Civil Engineering
Dr. John C. Falcocchio - Civil Engineering
Dr. Elena Prassas - Civil Engineering

Rensselaer Polytechnic Institute
Dr. José Holguín-Veras - Civil Engineering
Dr. William "Al" Wallace - Systems Engineering

Rochester Institute of Technology
Dr. James Winebrake - Science, Technology and Society/Public Policy
Dr. J. Scott Hawker - Software Engineering

Rowan University
Dr. Yusuf Mehta - Civil Engineering
Dr. Beena Sukumaran - Civil Engineering

State University of New York
Michael M. Fancher - Nanoscience
Dr. Catherine T. Lawson - City & Regional Planning
Dr. Adel W. Sadek - Transportation Systems Engineering
Dr. Shmuel Yahalom - Economics

Stevens Institute of Technology
Dr. Sophia Hassiotis - Civil Engineering
Dr. Thomas H. Wakeman III - Civil Engineering

Syracuse University
Dr. Riyad S. Aboutaha - Civil Engineering
Dr. O. Sam Salem - Construction Engineering and Management

The College of New Jersey
Dr. Thomas M. Brennan Jr - Civil Engineering

University of Puerto Rico - Mayagüez
Dr. Ismael Pagán-Trinidad - Civil Engineering
Dr. Didier M. Valdés-Díaz - Civil Engineering

UTRC Consortium Universities

The following universities/colleges are members of the UTRC consortium.

City University of New York (CUNY)
Clarkson University (Clarkson)
Columbia University (Columbia)
Cornell University (Cornell)
Hofstra University (Hofstra)
Manhattan College (MC)
New Jersey Institute of Technology (NJIT)
New York Institute of Technology (NYIT)
New York University (NYU)
Rensselaer Polytechnic Institute (RPI)
Rochester Institute of Technology (RIT)
Rowan University (Rowan)
State University of New York (SUNY)
Stevens Institute of Technology (Stevens)
Syracuse University (SU)
The College of New Jersey (TCNJ)
University of Puerto Rico - Mayagüez (UPRM)

UTRC Key Staff

Dr. Camille Kamga: *Director, UTRC*
Assistant Professor of Civil Engineering, CCNY

Dr. Robert E. Paaswell: *Director Emeritus of UTRC and Distin*guished Professor of Civil Engineering, The City College of New York

Dr. Ellen Thorson: *Senior Research Fellow*

Penny Eickemeyer: *Associate Director for Research, UTRC*

Dr. Alison Conway: *Associate Director for Education*

Nadia Aslam: *Assistant Director for Technology Transfer*

Dr. Wei Hao: *Post-doc/ Researcher*

Dr. Sandeep Mudigonda: *Postdoctoral Research Associate*

Nathalie Martinez: *Research Associate/Budget Analyst*

Tierra Fisher*: Office Assistant*

Andriy Blagay: *Graphic Intern*

TECHNICAL REPORT STANDARD TITLE PAGE

1. Report No.	2.Government Accession No.	3. Recipient's Catalog No.

4. Title and Subtitle Techniques for Efficient Detection of Rapid Weather Change and Analysis of their Impacts on a Highway Network	5. Report Date October 6, 2017 6. Performing Organization Code

7. Author(s) Catherine T. Lawson, University at Albany Feng Chen, University at Albany Adil Alim, University at Albany Joshi Aparna, University at Albany	8. Performing Organization Report No.

9. Performing Organization Name and Address University at Albany 1400 Washington Avenue Albany, New York 12222	10. Work Unit No. 11. Contract or Grant No. 49198-39-28

12. Sponsoring Agency Name and Address University Transportation Research Center The City College of New York 137th Street and Convent Ave, New York, NY 10031 Federal Highway Administration U. S. Department of Transportation Washington, D. C.	13. Type of Report and Period Covered Final Report 9/1/2016 – 8/31/2017 14. Sponsoring Agency Code

15. Supplementary Notes

16. Abstract

Adverse weather conditions have a significant impact on the safety, mobility, and efficiency of highway networks. Annually, 24 percent of all crashes, more than 7,400 roadway fatalities, and over 673,000 crash related injuries were caused by adverse weather conditions between 1995 and 2005 [1]. In addition, weather contributed to 23 percent of all non-reoccurring delay and approximately 544 million vehicle hours of delay each year [2]. Nearly 2.3 billion dollars each year are spent by transportation agencies for winter maintenance that contribute to close to 20 percent of most DOTs yearly budgets [2]. These safety and mobility factors make it important to develop new and more effective methods to address road conditions during adverse weather conditions.

This project develops techniques for efficiently detecting rapid weather change events and analyzing their impacts on the traffic flow characteristics of a highway network. It is composed of three components, including 1) detection of rapid weather change events in a highway network using the streaming weather information from a sensor network of weather stations; 2) detection of rapid change events on the traffic flow characteristics (e.g., travel time) of the highway network; and 3) analysis of correlations between the detected weather and traffic change events in space and time. The proposed approach was applied to a weather dataset provided by New York State Mesonet and a traffic flow dataset, the National Performance Management Research Data Set (NPMRDS), provided by NYSDOT, from Mar. 1, 2016 to Dec. 31, 2016. The empirical results provide potential evidence about the significant impacts of rapid weather change events on traffic flow characteristics of the Interstate 90 (I-90) Highway in the state of New York. The limitations of the proposed approach and the empirical study are also discussed.

17. Key Words	18. Distribution Statement

19. Security Classif (of this report)	20. Security Classif. (of this page)	21. No of Pages	22. Price
Unclassified	Unclassified		

Form DOT F 1700.7 (8-69)

Disclaimer

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. The contents do not necessarily reflect the official views or policies of the UTRC. This report does not constitute a standard, specification or regulation. This document is disseminated under the sponsorship of the Department of Transportation, University Transportation Centers Program, in the interest of information exchange. The U.S. Government assumes no liability for the contents or use thereof.

Contents

1 Introduction **1**

2 Literature Review **1**

3 Weather Station and Traffic Message Channel (TMC) real networks **3**

3.1 Mesonet Weather Station Network 3
3.2 Traffic Message Channel (TMC) Network 4
3.3 Data Description 6
3.3.1 Weather Data 6
3.3.2 Traffic Data 7

4 Methodology and Proposed Approach **8**

4.1 Problem definition 9
4.2 Weather change event Detection 11
4.2.1 Normalization of Time Window 14
4.2.2 Change event detection 15
4.2.3 Ranking of the results 16
4.3 Traffic change event Detection 16
4.4 Multi-Variable change event Detection 17
4.4.1 Selection of Related Variables 17
4.4.2 Multi-GraphMP algorithm 18
4.5 Correlation Study 19
4.5.1 Problem definition 19
4.5.2 Correlation Statistic Function 20
4.5.3 Hypothesis Testing 21

5 Experiments and Discussions **23**

5.1 Experiments 23
5.1.1 Time-Lag Effect and Detection 24
5.1.2 Weather Change Event Detection 25
5.1.3 Traffic Change Event Detection 27
5.1.4 Correlation Study 31
5.2 Discussions and Limitations 33

6 Conclusion **34**

1 Introduction

Analysis of the rapid (or sudden) weather changes has real-world implications in the transportation field such as improvement of transportation safety, mobility, and efficiency in response to rapid weather changes. Weather acts through visibility impairments, precipitation, high winds, and temperature extremes to affect driver capabilities, vehicle performance (i.e., traction, stability, and maneuverability), pavement friction, roadway infrastructure, crash risk, traffic flow, and agency productivity. On average, there are over 5,748,000 vehicle crashes each year. Approximately 22% of these crashes, nearly 1,259,000, are weatherrelated. Weatherrelated crashes are defined as those crashes that occur in adverse weather (i.e., rain, sleet, snow, fog, severe crosswinds, or blowing snow/sand/debris) or on slick pavement (i.e., wet pavement, snowy/slushy pavement, or icy pavement). On average, nearly 6,000 people are killed and over 445,000 people are injured in weather-related crashes each year [3]. These safety and mobility factors make it important to develop new and more effective methods to address road conditions during adverse weather conditions.

Existing techniques to detect a change point in weather conditions [4, 5, 6, 7, 8] either do not work with sensor networks or explore only regular-shaped subsets. The impact of weather conditions on traffic conditions has been studied in [9, 10, 11, 12, 13, 14, 15, 16], however, the impact of rapid weather changes on traffic conditions is not well-explored. Also, [17, 18, 19, 20] study the single point traffic flow forecasting, however, there is no multi-point traffic flow forecasting model that can deal with distribution changes. This project aims to address three main research questions:

- How can rapid weather changes in a highway network be detected in real time using the streaming multivariate weather information collected from weather stations near the network?
- How can rapid traffic flow changes in the highway network be detected in real time using the steaming traffic flow information collected from the highway network sensors?
- How can the significance of correlations between rapid weather changes and traffic flow changes in space and time be well assessed?

The remainder of this report is organized as follows: Section 2 reviews existing literature on rapid weather change detection. Section 3 proposes the modeling of real weather station network and Traffic Message Channel network into graphs and provides details about data processing. Section 4 introduces our research methodology. Section 5 presents the experiment and numerical results followed by the concluding remarks and future research directions in Section 6.

2 Literature Review

The literature review is organized based on the three research questions as discussed in the previous section.

1) **Detection of rapid weather changes.** This is also called a change point detection problem, and refers to the identification of abrupt variation of weather variables (e.g., temperature, humidity, wind, gust, pressure, and solar radiation) in a certain geographic region and a time point due to distributional or structural changes. A number of algorithms are available for change point detection, including detection of a change point in a univariate or multivariate time series collected from a single sensor node (weather station) [4, 5, 6], and detection of a change point in a collection of multivariate time series from multiple sensor nodes in a sensor network, where an unknown subset of sensor nodes are affected by the change point [7, 8]. As there are a large (exponential) number of possible subsets, much of this literature only explores regular-shaped subsets, such as circles and rectangles, in order to restrict the search space. To our knowledge, only a few references have addressed techniques for detecting a change point in a sensor network, where both *the subset of sensor nodes* and *the subset of weather variables* that are impacted by the change point are unknown [7, 8]. In particular, Neil explores space-time scan statistics to identify the best combination of subsets sensor nodes and weather variables as the indicator of a change point [7]. Jiang et al. explore joint sparse principle component analysis (PCA) algorithms to identify the indicating subsets of sensor nodes and variables for detecting a change point [8].

2) **Statistical impact analysis.** The impact of weather conditions (variables) on traffic conditions (e.g., speed, volume, travel time) was largely studied using statistical analysis techniques based on linear or logistic regressions [9], such as the impact of rainy weather on the speed variance of rural highways [10], the impact of cold and snow on traffic volumes [11], the impact of snow on travel time [12], and several others [13, 14]. Only a few references have studied the impact of weather conditions using nonlinear techniques [15, 16]. In particular, Mohammed et al. explore a mixture of linear regression models to study the impact of weather events (e.g., rain, fog, haze) on congestion identification [15]. Martchouk and Mannering apply a first-order autoregressive model to study the impact of weather events on travel time [16]. To our knowledge, there is limited work that has studied the impact of rapid weather changes on traffic conditions.

3) **Short-term traffic flow forecasting.** A number of algorithms are available for traffic flow forecasting [17, 18]. Most of the algorithms have focused on "one point" (or single road link) short-term traffic prediction that considered only the temporal domain and did not take into account the dependencies between road links (spatial domain) [18]. These so-called univariate methods are generally based on the use of time-series-based methods, such as the autoregressive integrated moving average (ARIMA) [19], back-propagation neutral networks, nonparametric regressions, and several others [20]. Despite these successes in single-point prediction, recent advances demonstrate that the multipoint forecasting models that take into consideration "geographic advantage" of urban network provide better prediction results, such as the spatio-temporal (ST) ARIMA [21], generalized STARIMA [22], and our recently proposed spatio-temporal random effects (STRE) model [18]. All the preceding forecasting mod-

els assume that the distribution of traffic conditions does not change over time. However, rapid weather changes will lead to change of distribution of traffic flow. To our knowledge, there is no multi-point traffic flow forecasting model that can deal with distribution changes. In the machine learning field, switching state-space models are designed to model change of distributions [23], but the integration of switching state-space models into existing traffic flow forecasting models is nontrivial.

Table 1 Abbreviations

Symbol	Meaning
T2M	Temperature that is taken at 2 meters from surfaceTemperature
T9M	Temperature that is taken at 9 meters from surface
TMC	Traffic Message Channel
I-90	Interstate 90
MAD	Mean Absolute Deviation
EMS	Evaluated Mean Scan
PIC	Pattern Instance Count
TS	Test Statistics
GLRT	Generalized Likelihood Ratio Test statistic

3 Weather Station and Traffic Message Channel (TMC) real networks

In this section, we discuss the weather and traffic flow datasets that we used during our experiments,and the modeling of the weather station netwrork graph and the TMC network graph. The weather dataset was originally provided by New York State Mesonet [24]. The traffic flow dataset was provided by the National Performance Management Research Data Set (NPMRDS) and has been made available by Albany Visualization and Informatics Lab (AVAIL) on behalf of New York State Department of Transportation (NYSDOT) and University Transportation Research Center (UTRC). In Section 3.1 and Section 3.2 we will discuss the details of the Mesonet Weather Station Network and Traffic Message Channel (TMC) Network, respectively.

3.1 Mesonet Weather Station Network

The New York State Mesonet consists of 125 stations across the state of New York. Each station houses a suite of automated sensors, which sample the sensors data every 3 to 30 seconds, then the data are packaged into 5-minute averages and transmitted in real-time to a central facility located at University at Albany (UAlbany). We selected 10 weather stations along the I-90 route, such as Batavia, South Bristol, Waterloo, Jordon, Westmorland, Central Square, Cold Brook, Sprakers, Cobleskill, Stephentown. Figure 1 shows locations of weather stations (represented by red stars) on the map.

Figure 1: Mesonet geographic locations

Based on the spatial neighborhood relationships of the selected weather stations, we modeled a graph viz. weather station graph, consisting weather stations as nodes in the graph and spatial neighborhood relationships between the nodes as edges. We indexed the 10 selected weather stations 0 to 9, {"BATA":0, "SBRI":1, "WATE":2, "JORD":3, "CSQR":4, "WEST":5, "COLD":6, "SPRA":7, "COBL":8, "STEP":9}, as shown in Figure 2.

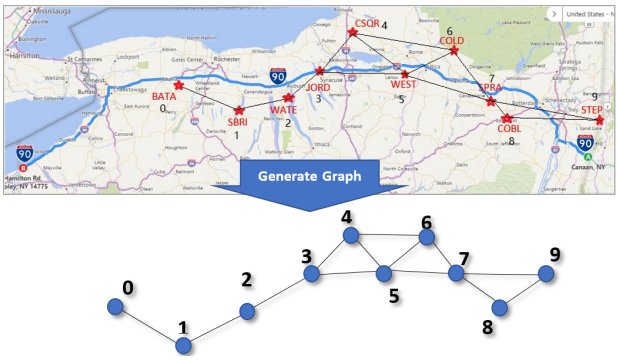


Figure 2: Weather station network modeling

3.2 Traffic Message Channel (TMC) Network

We have used the National Performance Management Research Data Set (NPMRDS). NPMRDS provides vehicle probe-based travel time data for passenger autos and trucks. The real-time probe data were collected from a variety of sources including mobile devices, connected autos, portable navigation devices, commercial fleet, and sensors. NPMRDS includes historical average travel times in 5-minutes increments on daily basis covering the National Highway System (NHS). The data are divided into two parts. The first part is a Traffic Message Channel (TMC) static file containing TMC information that does not change frequently. The second part includes travel times and identifies roadways geo-referenced to TMC location codes. The two datasets need to be combined in GIS-based software to provide the full picture. We used the data provided by NPMRDS which includes the traveling time of I-90 West and I-90 East routes. The I-90 West route section includes 52 TMCs and the I-90 East route section

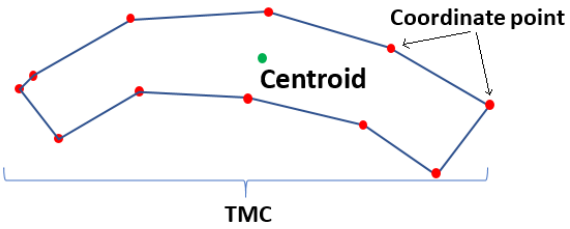


Figure 3: Single TMC centroid point calculation example

includes 54 TMCs. Each TMC has a length of the channel and geographic information of the channel where the geographic location coordinate is a polygon. We calculated the TMCs centroid geographic coordinates (i.e. latitude and longitude) and plotted them on the map. Figure 3 shows an example of the centroid point of a polygon latitude and longitude coordinates. Figure 4 is the map plot of I-90 West TMC centroid locations. The I-90 East TMCs map plot has the same shape as that of I-90 West and therefore is not shown.

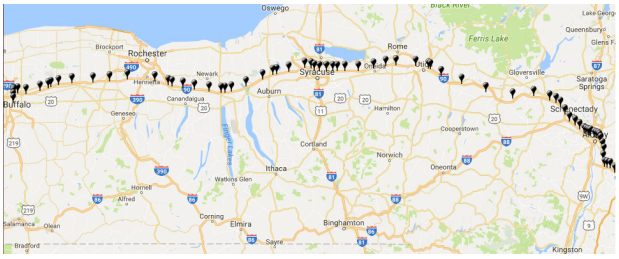


Figure 4: TMC centroid geographic locations of I-90 West route

We modeled the TMC networks for I-90 West route and I-90 East route into two TMC network graphs separately, based on neighborhood relationships of the TMCs. In each of the TMC network graphs the nodes represent TMCs and edges represent the neighborhood relationships between the TMCs. Because of the spatial structure of the TMC real network, we get two path graphs. Figure 5 shows the constructed graph of I-90 West (indexed the TMC with new ID 0 to 51) and I-90 East (indexed the TMC with new ID 0 to 53) TMC network.

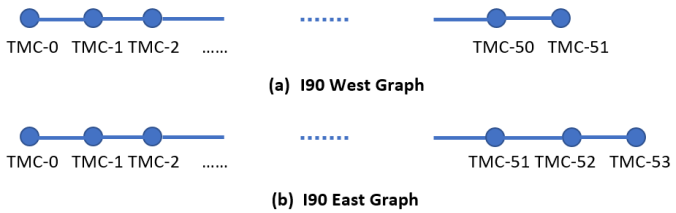


Figure 5: I-90 West and I-90 East TMC network graphs

3.3 Data Description

We have access to the weather and traffic data from Mar. 1, 2016 to Dec. 31, 2016, provided by Mesonet and NPMRDS. As discussed in Section 3.1 and Section 3.2, we modeled both weather station network and I-90 TMC networks as sensor graphs. The weather stations and TMCs both have records in 5-minutes increments. This means each weather station packages 5 minute averages of each weather variable and each TMC has the average traveling time for every 5 minutes. See Figure 6 for the examples of raw data (a) format of weather station record and (b) format of TMC record. All the data are in ".CSV" file format. Mesonet weather data has Station ID, datatime (consists of date and time which have 5 min time interval), 8 weather variables (T2M, T9M, rh, avg_wind_speed etc.). The variables temp_2m and temp_9m in 6 are represented by T2M and T9M throughout this report. NPMRDS data has TMC original ID, epoch (or the time slots, i.e. epoch number 1 represents time 00:00am, 2 represents 00:05am, etc.), npmrds_date, travelTime.

	station	datetime	temp_2m	temp_9m	rh [%]	avg_wind	max_wind	wind_direction	solar_radiati	station_press
2	BATA	20160601T000000	68.6	69	45	9.4	15.9	342	79	984.43
3	BATA	20160601T000500	68.4	68.8	44	11.4	15.5	336	63	984.47
4	BATA	20160601T001000	68.1	68.5	45	10.4	14.1	337	76	984.5
5	BATA	20160601T001500	67.9	68.2	46	9.1	12.3	331	61	984.57
6	BATA	20160601T002000	67.2	67.7	48	9	12.1	332	43	984.66
7	BATA	20160601T002500	66.9	67.4	48	8.1	12.6	333	30	984.71
8	BATA	20160601T003000	66.5	67.1	50	7.7	11.5	336	26	984.74
9	BATA	20160601T003500	66	67	51	6.8	9.6	338	16	984.8
10	BATA	20160601T004000	65.3	66.6	52	5.3	6.6			
	BATA	20160601T004500	65.3	66.6	53					

(a) Mesonet Raw Data Format

	tmc	epoch	npmrds_date	travelTime
2	104N04126	1	20160101	686
3	104N04126	2	20160101	687
4	104N04126	3	20160101	663
5	104N04126	4	20160101	644
6	104N04126	5	20160101	613
7	104N04126	7	20160101	661
8	104N04126	8	20160101	687
9	104N04126	9	20160101	674
10	104N04126	10	20160101	
	104N04126	11		

(b) NPRMDS Traffic Raw Data Format

Figure 6: Mesonet and NPMRDS raw data formats

For the weather and traffic data, we have records of 5 min intervals for all weather variables and TMC traveling time. We denote 5 min interval time as a time slot (epoch), in total we have of 288 time slots for each day. A Mesonet weather station records each weather variable every 5 minutes and obtains total 288 records per day. Similarly, each TMC has 288 traveling time records per day.

3.3.1 Weather Data

We obtained the weather data from Mar. 1, 2016 to Dec. 31, 2016 (a total of 306 days). For 10 selected stations, the complete data of the following 8 weather variables is available: T2M [degF], T9M [degF], Relative humidity [%], Average wind speed [mph], Max wind speed [mph], Wind direction [degree], Solar radiation $[W/m^2]$ and Station pressure [mbar].

Extraction of daily data We pre-processed the raw data and extracted the records of each variable. For each individual weather variable, we extracted the daily data for 10 selected weather stations. Further, we generated a file that includes 10 stations daily data. Figure 7 shows the weather variable T2M's values for 10 stations on Mar. 1, 2016. The 1st column is the weather station ID, the following 288 columns are the records of 288 time slots. The $cell(i, j)$ represents the T2M values of ith weather station at jth time slot on Mar. 1, 2016. For all 8 weather variables, we followed the same process. At the end, for each variable we got 306 files for 306 days and in total we had 8×306 i.e. 2448 files.

	Station_ID	slot-1	slot-2	slot-3	slot-4	slot-5	slot-6	slot-7	
2	0	38	37.9	38	35.6	31.5	30.9	30.5	...
3	1	42.7	42.6	42.7	42.7	42.6	42.5	42.7	...
4	2	44.9	45.1	44.7	45.1	45.3	45.3	45.3	...
5	3	44.4	44.4	44.5	44.6	44.7	44.6	44.6	...
6	4	41.5	41.9	41.9	41.6	41.9	42	42.3	...
7	5	42.5	42.4	42.5	42.4	42.2	42.2	42.1	...
8	6	33.5	34	34.5	35.3	35.5	35.8	35.8	...
9	7	34.3	34.1	34.2	34.4	34.5	34.4	34.1	...
10	8	41.9	42	42.3	42.9	43.3	43.2	43.4	...
11	9	29.6	29.9	29.5	29.2	29.2	28.8	28.8	...

Figure 7: Mar. 1, 2016 processed data for T2M

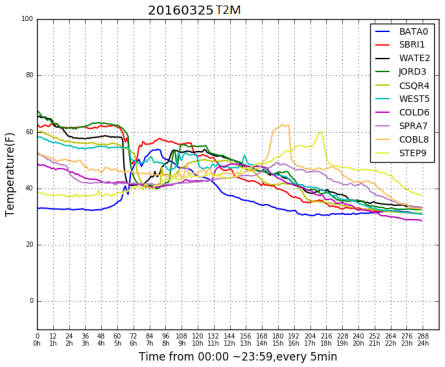
Example Plots Each weather variables have different shapes of plots. To show the visualization of our weather station network one day data, we give the example plots in this section. In Figure 8 and 9 we show the example plots of all 8 variables on Mar. 25, 2016.

3.3.2 Traffic Data

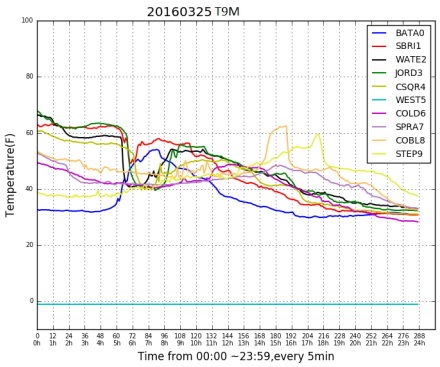
We have access to the traffic data from Mar. 1, 2016 to Dec. 31, 2016 (a total of 306 days same as the weather data) for both I-90 East and I-90 West route. For both, I-90 West and I-90 East route, we have the traffic traveling time data.

Extraction of daily data We pre-processed the raw traffic data in a similar way to the weather raw data as explained earlier. For all TMCs in I-90 West route, we extracted the average traveling time data, and we packaged it into a file that includes I-90 West route TMC daily data. Figure 10 shows the average traveling time for 52 TMCs on Mar. 1, 2016. The 1st column is the TMC new ID, the following 288 columns are the records of 288 time slots. The cell (i, j) represents the average traveling time values of ith TMC at jth time slot on Mar. 1, 2016. For TMCs in I-90 East route data, we followed the same process. At the end, for each route, we got 306 files for 306 days and in total, we had 2×306 i.e. 612 files.

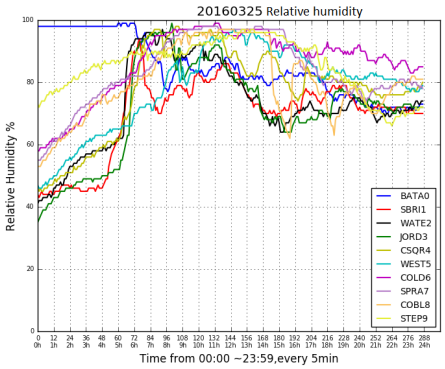
Example Plots In this section, we show an example plot for traveling time. The following plot is an example of traveling time plot for I-90 West route on Mar. 25, 2016 (see Figure 11). The x axis represent the time slots and the y axis



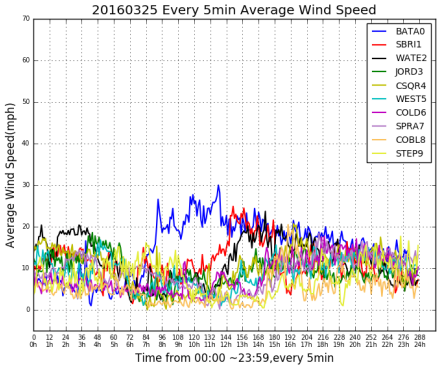
(a) T2M [degF]



(b) T9M [degF]



(c) Relative humidity [%]



(d) Average wind speed [mph]

Figure 8: Weather data example plots

represent the Traveling Time values for each time slot. Each different colored line represent the values of different TMCs.

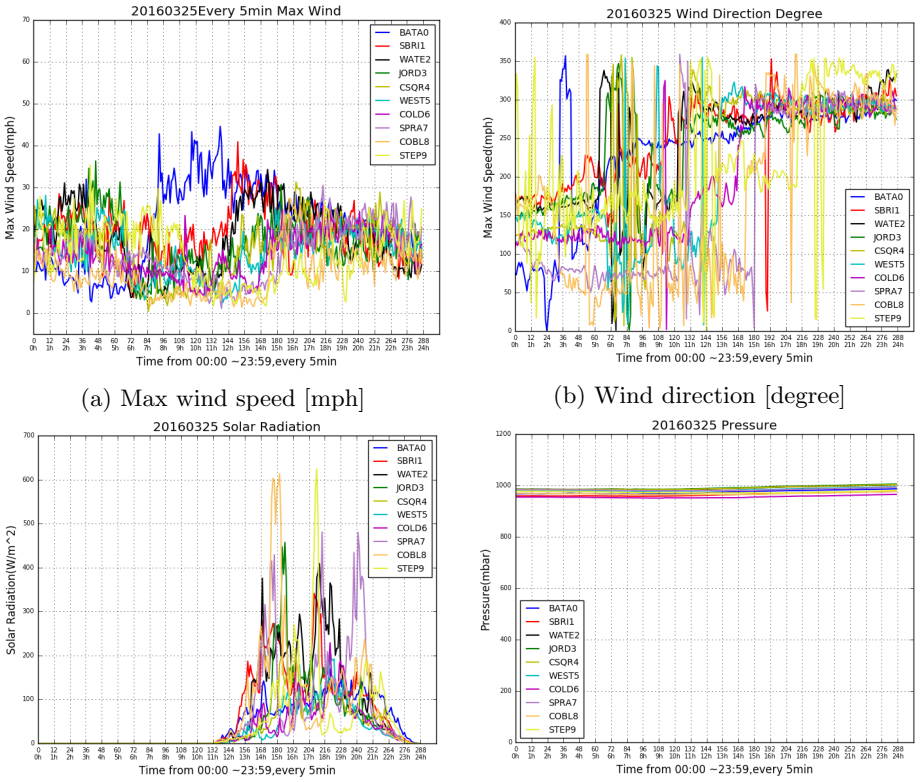
4 Methodology and Proposed Approach

In this section, we will explain our proposed approach which consists of three main components (See Figure 12):

1. **Weather change event detection.** Detection of change events from multiple weather variables data.

2. **Traffic change event detection.** Detection of change events from single traffic variable data.

3. **Correlation study between rapid weather change and rapid traffic change.** Study of the correlation between weather change events and traffic change events that are the result of our weather and traffic change event detection.



(a) Max wind speed [mph]

(b) Wind direction [degree]

(c) Solar radiation [W/m^2]

(d) Station pressure [mbar]

Figure 9: Weather data example plots

	TMC_ID	slot-1	slot-2	slot-3	slot-4	slot-5	slot-6	slot-7	slot-8	slot-9	slot-10
2	0	217	220	235	226	227	234	232	234	234	234
3	1	62	62	62	65	71	66	71	69	72	70
4	2	311	276	252	286	279	280	256	285	284	317
...	...	...	...	...	...	...	...	...	...	...	...
51	49	413	405	413	415	383	389	425	454	422	423
52	50	285	291	286	289	298	275	273	286	315	298
53	51	142	142	150	220	544	544	544	180	126	174

Figure 10: Mar. 1, 2016 processed data for Traveling Time

The main goal of this work is to develop efficient techniques to detect rapid weather changes in a highway road network using the streaming weather information from a sensor network of weather stations. We also developed similar techniques to detect (predict) traffic changes in TMC network. Further, we studied the correlation between the weather and traffic change events.

4.1 Problem definition

Both weather station network and TMC network are time evolving dynamic networks. Our main problem is to detect change events. A change event is represented by a tuple of 1) a subset of nodes (weather stations or TMCs); 2) a subset of variables and 3) a time interval (a continuous sequence of time slots) from the weather and traffic data. In this section, we will discuss the

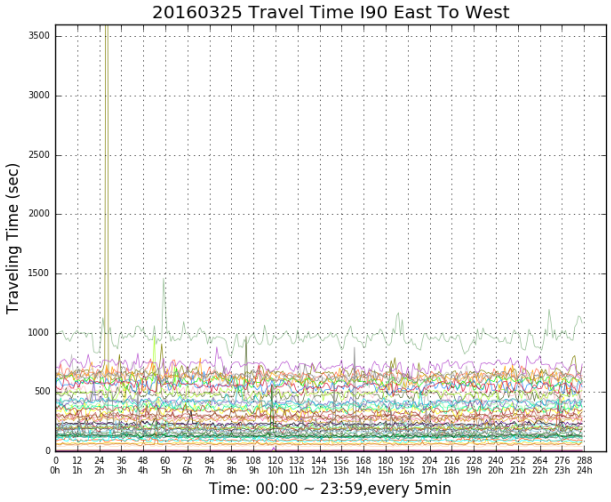


Figure 11: Mar. 1, 2016 I-90 West Traveling Time plot

weather station network and the weather change event detection problem. Due to the similarity of the problems and networks, the traffic change event detection problem is analogous to the weather change event detection problem.

The sensor network is defined as $\mathbb{G} = (\mathbb{V}, \mathbb{E}, x)$ where $\mathbb{V} = \{1, 2, ..., n\}$ refers to the set of sensor nodes (weather stations), $\mathbb{E} \subseteq \mathbb{V} \times \mathbb{V}$ refers to the set of edges, and $x(v)$ is a mapping function: $\mathbb{V} \to \mathbb{R}^{d \times T}$ that returns a matrix storing the measurements of d weather variables at node v for the time interval T (a continuous time slots). The objective is to identify a **change event** characterized as a tuple $(o, \mathcal{S}, \mathcal{R})$ where o is a time window, $\mathcal{S} \subseteq \mathbb{V}$ is a subset of stations and $\mathcal{R} \subseteq \{1, ...d\}$ is a subset of variables, where ($o = [o_{start}, o_{end}]$,o_{start} and o_{end} is the starting and ending time slots of time window o). The detection problem can be modeled using a hypothesis testing framework:

- **Null hypothesis** H_0 (there is no change event): $\{\mathbf{x}(i) \mid i \in \mathbb{V}\} \sim \mathcal{D}_0(\theta_0)$, where $\mathcal{D}_0$ refers to the distribution under the null hypothesis, and θ_0 is the parameter of this distribution.

- **Alternative hypothesis** $H_1(o, \mathcal{S}, \mathcal{R})$ (there is a change event): $\{x_i^t(v) \mid v \in \mathcal{S}, t \geq o_{start}$ and $t \leqslant o_{end}, i \in \mathcal{R}\} \sim \mathcal{D}_1(\theta_1)$, where $\mathcal{D}_1$ refers to a new (or changed) distribution corresponding to the change event within a given time window; $\{x_i^t(v) \mid v \notin \mathcal{S}$ or $t < o_{start}$ or $i \notin \mathcal{R}\} \sim \mathcal{D}_0(\theta_0)$, where $t < o_{start}$. It means that the values in the time window follow $\mathcal{D}_1(\theta_0)$ distribution and the historical values before this time window o follow $\mathcal{D}_0(\theta_1)$ distribution.

The change event detection problem can then be implemented in three following steps:

- **Step 1:** Application of Gaussian distribution to define $\mathcal{D}_0$ and $\mathcal{D}_1$ for the hypothesis testing framework.

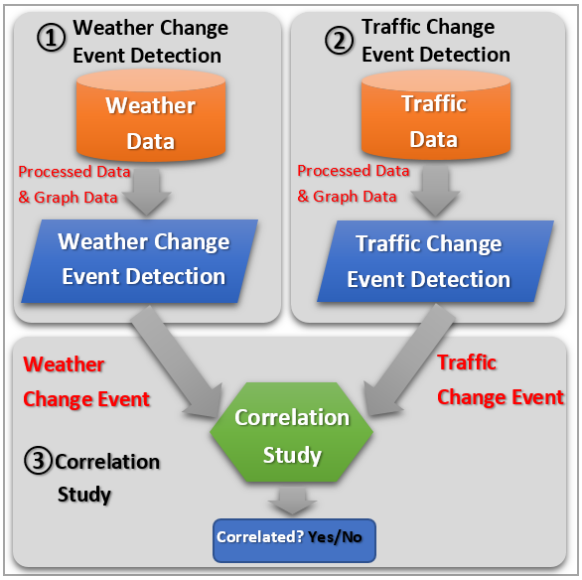


Figure 12: Project architecture

- **Step 2:** Identification of the best possible change event $(o, \mathcal{S}, \mathcal{R})$ by solving the optimization problem:

$$(\hat{o}, \hat{\mathcal{S}}, \hat{\mathcal{R}}) = \arg\max_{o,\mathcal{S},\mathcal{R}} F(o, \mathcal{S}, \mathcal{R}) = \frac{\text{Prob}(\textbf{data} \mid H_1(o, \mathcal{S}, \mathcal{R}))}{\text{Prob}(\textbf{data} \mid H_0)}, \tag{1}$$

  where $\textbf{data} = \{\mathbf{x}(i) \mid i \in \mathbb{V}\}$. The numerator in Equation 1, refers to likelihood of **data** under alternative hypothesis whereas, the denominator refers to likelihood of **data** under null hypothesis. The ratio $F(o, \mathcal{S}, \mathcal{R})$ is called generalized likelihood ratio test statistic (GLRT).

- **Step 3:** Estimation of the empirical p-value of estimated GLRT score: $F(\hat{o}, \hat{\mathcal{S}}, \hat{\mathcal{R}})$ using bootstrap methods. If the empirical p-value is below a predefined significance level (e.g., $\alpha = 0.05$) then $(\hat{o}, \hat{\mathcal{S}}, \hat{\mathcal{R}})$ is returned as a significant change event; otherwise, no change event is returned.

In our recent work, we have developed several efficient algorithms for optimizing a general GLRT function in nearly-linear time in a univariate sensor network ($d = 1$) [25, 26]. We will extend this work and design efficient algorithms to optimize Problem (1) with $d > 1$.

4.2 Weather change event Detection

In previous Section 4.1, we discussed the formulation of our change event detection problem. In this section, we will introduce our proposed algorithms to detect the multi-variable weather change events. A weather change event is a tuple of **a connected subset of weather station nodes**, **a subset of rapid change related weather variables** and **a time interval** in which the rapid change have occurred. We designed an invariant of Graph-MP method [25]

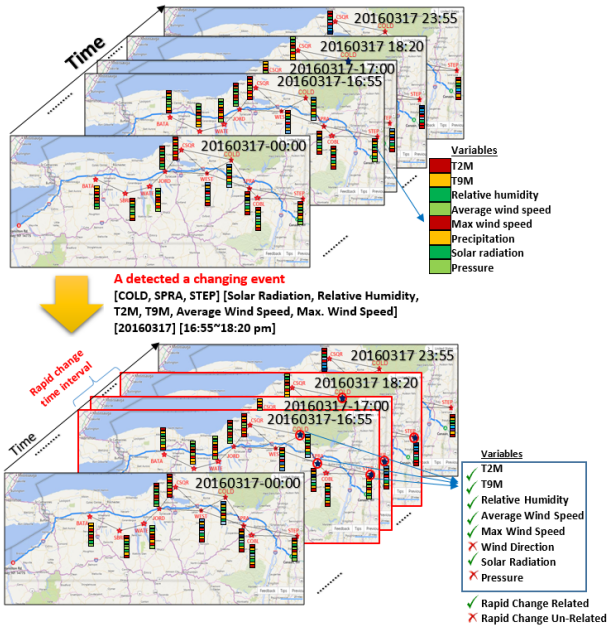


Figure 13: An example of weather change event detection

to address this problem in order to achieve a nearly-linear time complexity and high accuracy. In Figure 13, the top sub-figure shows a time evolving attributed graph for Mar. 17, 2016, in which each node is associated with multiple attributes. The bottom sub-figure of Figure 13 shows a detected weather change event. Figure 13 shows an illustrative example of weather change event detection which indicates that the time interval 16:55 to 18:20 on Mar. 17, 2016 at weather stations COLD, SPRA and STEP rapid change occurred and involved variables such as Solar Radiation, Relative Humidity, T2M, T9M, Average Wind Speed, Max. Wind Speed.

Before explaining the details of the change event detection algorithm, we would like to introduce several key notations:

- **Time Window** is a time interval that consists of continuous time slots and has minimum and maximum length. See Figure 14 (a).
- **Rapid weather change** is defined as the weather variables of weather station(s) that are rapidly decreasing or increasing in a given time window.
- **Time Window Value** (μ_{window}) is the mean value of data within the time window.
- **Historical Base Value** (μ_{base}) is the mean value of data over a given number of time slots in past (just before the current time window).
- **Changing value** ($\nabla\mu_{change}$) represents the normalized value of the given time window. The changing value is the difference between the mean of

the historical base and the mean of the time window, and it indicates the changing level of the values in the time window, $\nabla\mu_{change} = |\mu_{window} - \mu_{base}|$. See Figure 14 (b).

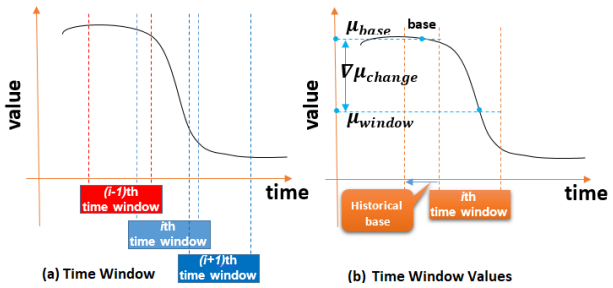


Figure 14: Time window conception

Algorithm 1 SINGLE-VARIABLE CHANGE EVENT DETECTION

Input: $D = \{D_1, \cdots D_{Num\_days}\}$
A,s, min_widnow,max_widnow, $hist\_base\_length$;
Output: List of change events $change\_event\_list$;
$change\_events \leftarrow \emptyset$;
$time_windows \leftarrow$getAllPossibleTimeWindow(min_window, max_window);
for $j = 1$ to Num_Days **do**
  for $k = 1$ to $time\_windows.Size$ **do**
    $hist\_base \leftarrow$ CalculateHistoricalBase($D_j, time\_windows[k]$, $hist\_base\_length$);
    $X \leftarrow$ NormalizeWindow($D_j, time\_windows[k]$);
    $\hat{X} \leftarrow |hist\_base - X|$
    $result \leftarrow$ GRAPHMP($A, \hat{X}, s, EMS$);
    $change\_event\_list$.add($result.Score, result.SubsetNodes$, $time\_windows[k], j$);
  end for
end for
$cchange\_event\_list \leftarrow$ RankBasedOnScore($change\_event\_list$)
return $change\_event\_list$;

Algorithm 1 is our proposed single variable change event detection algorithm. Algorithm 1 includes three main steps: 1) Normalization and calculation of changing the value of a given time window; 2) Detection of the most anomalous subset of the station(s) in a given time window and 3) Ranking the change event results.

In Algorithm 1, the input data $D = \{D_1, \cdots, D_{Num\_days}\}$ includes daily data of a single weather variable for whole weather station network. A is the adjacency matrix, s is the upper bound of the maximum subset of nodes (our algorithm will return at most 4*s nodes), min_window and max_window are minimum time window size and maximum time window size respectively, $hist\_base\_length$ is the size of historical base time window.

4.2.1 Normalization of Time Window

In Algorithm 1, at **Line 5**, all possible time windows for given minimum and maximum size of the time window are generated. Outer loop at **Line 6** loads the daily data one by one. Inner loop at **Line 7** iterates over all possible time windows to test whether there exists a change event in any of these time windows. At **Line 8**, calculates the historical base of the current time window, which is mean of the values of previous time slots length with $hist\_base\_length$, is calculated. By default, we have used last one hour (12 time slots just before the current time window) data's mean value as the historical base. At **Line 9∼10**, the values in current time window are normalized by calculating the absolute difference of the mean values of current time window and historical base of each node.

For normalization at each weather station node, we normalized the values

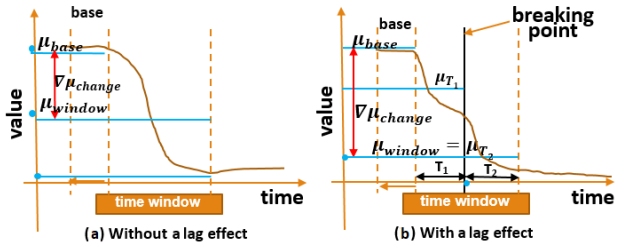


Figure 15: Cases of with and without Lag effect

in the given time window into a single changing value. To eliminate the lag effect within the time window, we considered two cases when we calculated the changing value:

- **Case 1: Without a lag effect in the time window**. In this case no obvious changing point exists in the current time window. It means that, there is no lag effect, and the changing value of the given time window is equal to $\nabla\mu_{change} = |\mu_{window} - \mu_{base}|$. See Figure 15 (a).
- **Case 2: With a lag effect in the time window**. If there is a lag effect in the time window, the mean value of the time window cannot represent the true changing level of the window. To decrease the impact of the lag effect to the value of time window, we need to find a breaking point that splits the time window into two sub-windows viz. T_1 and T_2 (See Figure 15 (b)), such that, their mean values are μ_{T_1} and μ_{T_2}. This breaking point guarantees that the difference of μ_{T_1} and μ_{T_2} is the maximum difference compared to the difference between the mean values in sub-windows generated by any other breaking points. The changing value of the given window is $\nabla\mu_{change} = \max\{|\mu_{T_1} - \mu_{base}|, |\mu_{T_2} - \mu_{base}|\}$.

Figure 16 shows a visualization example of the normalization step on the temperature data for a given time window. In this example, after normalizing the given time window, each weather station node has a changing value, which represents the changing level of the weather station in the given time window.

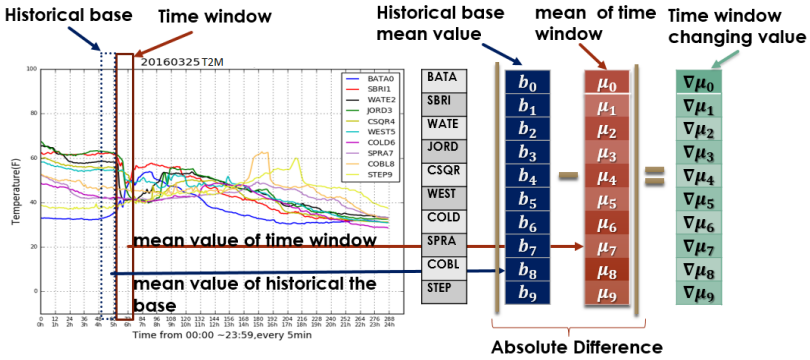


Figure 16: Time window value normalization for each node

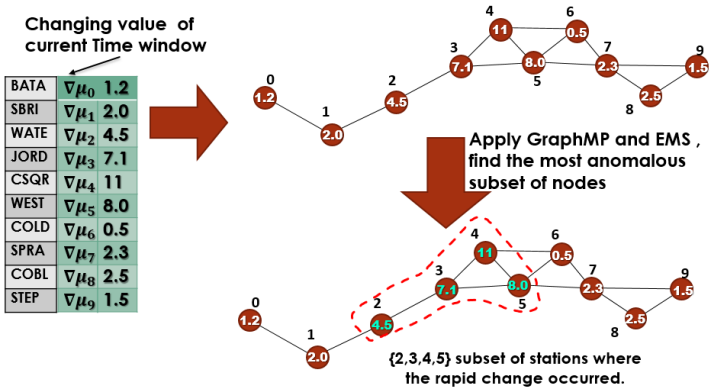


Figure 17: An example of most anomalous subset of station(s) detection

4.2.2 Change event detection

After normalization of the time window data, each weather station node has a changing value. Our problem is to find the most anomalous, connected subset of weather stations such that, the summation of the changing values of these weather station nodes is the maximum. To solve this problem we applied the Evaluated Mean Scan Statistics (EMS) as the score function and Graph-structured matching pursuit (Graph-MP) [25] method to detect a subset of stations that are spatially connected and their weather variables changed rapidly within the time window.

- **Evaluated Mean Scan (EMS) Statistics**. In EMS statistics, we assume there is a ground set $V = \{x_i\}_1^n$, x_i is following Normal distribution and S is some unknown anomalous cluster, where $S \subseteq V$. The aim is to decide between the null hypothesis $H_0 : x_i\ \mathbb{N}(0,1), \forall i \in V$ and the alternative $H_1 : x_i\ \mathbb{N}(\mu,1), \forall i \in S$ and $x_i\ \mathbb{N}(0,1), \forall i \in V - S$. So the EMS scan statistics for a cluster S is:

$$EMS(S) = \frac{1}{\sqrt{|S|}} \sum_{i \in S} x_i. \tag{2}$$

Under some conditions, the test of rejecting H_0 for larger values of $\max_{S \subseteq V}$: $EMS(S)$ is statistically optimal.

- **Graph-MP method**. Graph-MP is an efficient approximation algorithm. Graph-MP can be applied to optimize a variety of graph scan statistic models for the task of connected subgraph detection. We applied Graph-MP algorithm to solve our problem. Graph-MP optimizes our objective EMS score function and returns a subset of weather station nodes that are connected with each other.

In Algorithm 1, at **Line 11**, Graph-MP takes the network graph adjacency matrix A, normalized changing values of each node $\hat{X}$, sparsity constraint s and our main objective EMS score function as the input. The Graph-MP algorithm will return a subset of connected nodes and their objective EMS score. If the returned subset of nodes has a larger change level (overall changing amount of the variable in the given time window) then its corresponding EMS score will be larger, otherwise, the score will be smaller.

4.2.3 Ranking of the results

Ranking the change events. After running Algorithm 1 for all the data and for all possible time windows, we obtained a list of candidate change events with their corresponding EMS scores. In **Line 15**, the candidate change events are ranked based on their EMS scores. The candidate change events with higher EMS scores are most likely to be the true rapid change events.

Filtering the results. In the list of ranked candidate change events, if multiple events overlap (if the time window of the events overlapped with each other equal or greater than 80%) with each other on time intervals and have weather station nodes in common, then only the event/s with the highest EMS score among the overlapping events are kept.

4.3 Traffic change event Detection

We have two TMC network graphs for I-90 West and I-90 East (See Figure 5). Both of these networks have a similar shape and each TMC has average traveling time as an attribute. Since the Traffic change event detection is similar to the single variable weather change event detection method, we can use Algorithm 1 to find all change events.

Compared to the single variable weather change event detection, the traffic change event has different algorithm parameter settings and different time window normalization step. We will discuss the parameter setting in Experiments Section 5. Since the traveling time data has special change event shape, we assume all time windows without lag effect. Currently, we only have the data for traveling time available. If in the future, more traffic variables data become available, then we can apply Algorithm 2 viz. Multi-variable change event detection algorithm.

Algorithm 2 MULTI-VARIABLE CHANGE EVENT DETECTION

1: **Input**: $D = \{D_{1,1}, \cdots, D_{1,Num\_days}, D_{2,1} \cdots, D_{Num\_Vars,Num\_Days}\}$
2: A,s, min_widnow,max_widnow, $hist\_base\_length$, $threshold$;
3: **Output**: List of change events $change\_event\_list$;
4: $change\_event\_list \leftarrow \emptyset$;
5: $time\_windows \leftarrow$ getAllPossibleTimeWindow(min_window, max_window);
6: **for** $j = 1$ to Num_Days **do**
7: **for** $k = 1$ to $time\_windows.Size$ **do**
8: $related\_vars \leftarrow \emptyset$;
9: $\hat{X}_{related} \leftarrow \emptyset$;
10: **for** $i = 1$ to Num_var **do**
11: $hist\_base \leftarrow$ CalculateHistoricalBase($D_{i,j}$, $time\_windows[k]$, $hist\_base\_length$);
12: $X_i \leftarrow$ NormalizeWindow($D_{i,j}, time\_windows[k]$);
13: $\hat{X}_i \leftarrow |hist\_base - X_i|$
14: $single\_var\_result \leftarrow$ GRAPHMP($A, \hat{X}_i, s, EMS$);
15: **if** $single\_var\_result.Score > threshold[i]$ **then**
16: $related\_vars$.add(i);
17: $\hat{X}_{related}$.add($\hat{X}_i$);
18: **end if**
19: **end for**
20: $result \leftarrow$ MULTIGRAPHMP($A, \hat{X}_{related}, s, related\_vars$, $time\_windows[k], MultiEMS$);
21: $change\_event\_list$.add($result.Score$, $result.SubsetNodes, related\_vars, time\_windows[k], j$);
22: **end for**
23: **end for**
24: $change\_event\_list \leftarrow$ RankBasedOnScore($change\_event\_list$)
25: **return** $change\_event\_list$;

4.4 Multi-Variable change event Detection

In Section 4.2, we explained the single variable change event algorithm. In this section, we will explain the multi-variable change event detection algorithm. In the weather station network, each node is associated with multiple weather variables, and we use Algorithm 2 to detect multiple variables related to weather change event. Compared to Algorithm 1, Algorithm 2 has two main differences: **1) Selection of the subset of change event related variables;** and **2) Multi-GraphMP algorithm**.

4.4.1 Selection of Related Variables

In Algorithm 2, the inner loop **Line 10∼19**, iterates over all the variables in the given time window and separately runs the Graph-MP method for each variable. In **Line 11** $D_{i,j}$ indicates the jth day data matrix of ith variable for weather station network graph. **Line 15∼18** compare the EMS score of each variable change event with their threshold EMS score. In a given time window if the

EMS score of a variable change event is higher than the corresponding variable EMS threshold score, then we assume this variable has a rapid change in the current time window. **Line 17** stores the changing values of ith variable for current time window to the list $\hat{X}_{related}$.

Weather Variables Anomalous Threshold Score For each weather variable data (for all available data from Mar. 1, 2016 $\sim$ Dec. 31, 2016) we run Algorithm 1 separately and got a change event list for each variable. We calculated the *median*, *mad* (mean absolute deviation), *min* and *max* EMS scores of each variable change event list separately. We set $\boldsymbol{threshold(var_i) = median_i + 2 * mad_i}$ as the threshold score for each variable, where i indicates the index of the variable, $i = 0 \sim 7$. For a given time window, if the EMS score of detected change event is equal to or greater than the threshold EMS score, then this change event is most likely to be the true rapid change event, as we call it change event related variable. Table 2 shows the threshold of each weather variable.

Table 2 EMS Threshold Score of Weather Variables

Type	**median+2*mad**	**median**	**mad**	**min**	**max**
T2M	12.52	5.78	3.37	1.01	30.48
T9M	11.47	5.43	3.02	1.02	29.98
Pressure	3.28	1.78	0.75	1.0	9.14
Avg. Wind Speed	10.6	7.28	1.66	5.5	29.45
Wind Direction	510.36	370.46	69.95	188.17	663.72
Max. Wind Speed	17.18	11.3	2.94	8.0	44.71
Relative Humidity	35.82	18.14	8.84	1.82	77.6
Solar radiation	888.87	518.03	185.42	1.0	1418.56

4.4.2 Multi-GraphMP algorithm

Algorithm 2, called Multi-Variable change event Detection (Multi-GraphMP) Algorithm, is a variant of GraphMP algorithm. At **Line 20** MultiGraphMP takes A, the network adjacency matrix, $\hat{X}_{related}$, the changing values of related variables in the current time window, $related\_vars$, an indices list of change event related weather variables and $MultiEMS$, a variant of EMS score function as input. The main difference between GraphMP and Multi-GraphMP is that in Multi-GraphMP the EMS score is summation of the EMS score related variables in a given set of nodes:

$$MultiEMS(S, R) = \sum_{j \in R} EMS(S, j) = \sum_{j \in R} \sum_{i \in S} \frac{1}{\sqrt{|S|}} x_{i,j}, \tag{3}$$

where R is the list of variable indices and x corresponds to the variable $\hat{X}_{related}$, $x_{i,j}$ represents the changing value of ith node jth variable in current time window. **Line 20** MultiGraphMP returns the subset of nodes and its Multi-EMS score. **Line 21** adds a new detected changing result to the result list. $result.Score$ is the MultiEMS socre, $result.SubsetNodes$ is the change event

weather station node(s), *related_vars* is a list indices of related variables, j is the jth day and *time_window*$[k]$ is the time slot interval of the current time window.

4.5 Correlation Study

Our another important task is to study the correlation between weather change event and traffic change event. **Is there any correlation between the weather change events and the traffic change events?** In this section, we will introduce **Test Statistic Function** and **Hypothesis Testing Method**.

4.5.1 Problem definition

In Section 4.1, we defined the **change event** as a tuple (a subset of nodes, a subset of variables, time slot). To study the correlation of weather and traffic event, we generate the weather/traffic **Event** by splitting a weather/traffic change event individual weather/traffic events. A general event is defined as:

$$[\boldsymbol{Event\ Type}][\boldsymbol{Event\ Location}][\boldsymbol{Event\ Time}].$$

The weather event is defined as:

$$(\mathbf{0}, \boldsymbol{Weather\ Station\ Location}, \boldsymbol{Occur\ Time}),$$

where "0" indicates the weather event type, the weather station location is the latitude and longitude of weather stations. Event occur time consists of date and time slot. e.g. we have a weather event (0, (43.02,78.34), 20160301230). The occur time "20160301230" first 8 digits indicate the date Mar. 1, 2016 and last 3 digits indicate the time slot 230 (i.e. 230 represents 19:10pm).

The traffic event is defined as:

$$(\mathbf{1}, \boldsymbol{TMC\ Location}, \boldsymbol{Occur\ Time}),$$

where "1" indicates traffic event type, TMC location is the latitude and longitude of TMC. Event occur time consists of date and time slot. e.g. (1,(44.22,76.18),20160301232).The occur time "20160301232" first 8 digits indicates the date (Mar. 1, 2016) and last 3 digits indicates the time slot 232 (i.e. 232 represents 19:20pm).

We have given the definition of two types of events, and both types of events have event type, event location and event occur time. Now we will give an example that shows how we generate the events from the change events. For example, we have a weather change event ([2,3,4],[6,7][120,121,122],20160505), where [2,3,4] is a list of weather station indices, [6,7] is a index list of related variables and [120,121,122] are continuous time slots, 20160505 is the date. To generate weather events, we will split the weather change event into events. After splitting above given example weather change event, we get a list of weather events (in totally 9 weather events): (0, 2, 20160505120), (0, 2, 20160505121), (0, 2, 20160505122), (0, 3, 20160505120), $\cdots$, (0, 4, 20160505122).

Assume, we have n weather events and m traffic events. $\mathbb{W} = \{e_1^w, \cdots, e_n^w\}$ is a set of weather events, and $\mathbb{T} = \{e_1^t, \cdots, e_m^t\}$ is a set of traffic events, where $e_i^w = (0, e_i^w.location, e_i^w.time)$ and $e_i^t = (1, e_i^t.location, e_i^t.time)$. We have only two types of events, the weather events and traffic events. Now we can define our hypothesis:

- **Null Hypothesis(H_0)** The events of these two types are distributed in the space and time independently.
- **Alternative Hypothesis(H_1)** The events of these two types are distributed in the space and time dependently.

In the following two sections we will discuss the design of a statistical model to test our hypothesis.

4.5.2 Correlation Statistic Function

We have two types of events, and these events only can occur at a specific location, like the weather station locations or TMCs locations. Currently, we have used Pattern Instance Count (PIC) [27], the number of pairs of events (instance) of two different types within a given radius r.

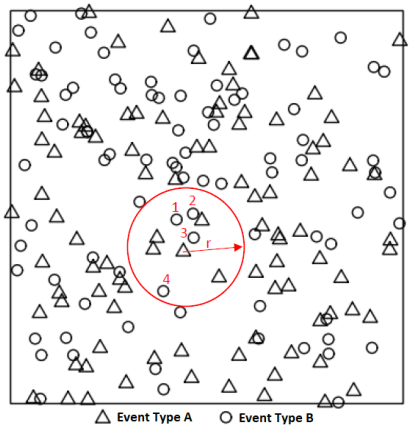


Figure 18: An example of PIC score calculation

In Figure 18, assume there are two types of events A and B spatially distributed in the study region, and each event has its location information. To calculate the PIC score for a set of events (including A and B), for each Type A events we count the Type B events within the given radius r and the final PIC score is the summation of all pair counts. The PIC(A_Events, B_Events, r) score function has three inputs, where A_Events is a set of A type of events and B_Events is a set of B type of events, and r is the testing radius. PIC score is the number of pairs of events of two different types (e.g. Type A and B) within a given radius r. In Figure 18 the radius of the red circle is r, and its center is a Type A event (triangle symbol). To calculate the PIC score of this single event, we need to count the Type B events within the circle, which is 4.

Algorithm 3 Original PIC Algorithm

```
1: Input: 𝕎 //set of weather events;
2:   𝕋 //set of traffic event
3:   r //testing geographic radius
4: Output: PIC score;
5: PIC = 0.0 ;
6: for i = 1 to 𝕎.Size do
7:     for j = 1 to 𝕋.Size do
8:         if Distance(𝕎[i], 𝕋[j]) <= r then
9:             PIC ← PIC + 1.0;
10:        end if
11:    end for
12: end for
13: return PIC;
```

Algorithm 4 PIC Spatio-Temporal Algorithm

```
1: Input: 𝕎 //set of weather events;
2:   𝕋; //set of traffic event
3:   r_g; //testing geographic radius
4:   r_t; //testing time radius
5: Output: PIC score;
6: PIC = 0.0 ;
7: for i = 1 to 𝕎.Size do
8:     for j = 1 to 𝕋.Size do
9:         if Distance(𝕎[i], 𝕋[j]) <= r_g and |𝕎[i].time − 𝕋[j].time| <= r_t then
10:            PIC ← PIC + 1.0;
11:        end if
12:    end for
13: end for
14: return PIC;
```

Algorithm 3 is the original algorithm to calculate PIC score for two spatially distributed the event types. In this time evolving network graph each event has the event geographic location and the event occur time. In our study, we treated event occur time as another "spatial coordinate". If the geographic distance of a pair of events is less than the given testing radius and the two events occur time interval is less than the given time threshold (or the time radius), then we can count this pair of events in the PIC score, otherwise, we can not count it in. Algorithm 4 is the revised version Algorithm 3. In **Line 9** we added the time threshold, such that, the time interval of the weather event and traffic event must be less than the time radius r_t.

4.5.3 Hypothesis Testing

To test our null hypothesis we need a test statistic that will have different values under the null hypothesis and the alternatives we care about. We then need to compute the sampling distribution of the test statistic when the null hypothesis is true. For some test statistics and some null hypotheses, this can be done analytically. The p-value is the probability that the test statistic would be at least

as extreme as we observed if the null hypothesis is true. A **permutation test** gives a simple way to compute the sampling distribution for any test statistic, under the strong null hypothesis that a set of genetic variants has absolutely no effect on the outcome.

Random Permeation Test is testing the Test Statistics (TS) of the original data and the TS of randomly permuted data. Random permutation hypothesis:

- **Null Hypothesis(H_0)** Random permutation has no effect to Test Statistics.
- **Alternative Hypothesis(H_1)** Random permutation changes Test Statistics results.

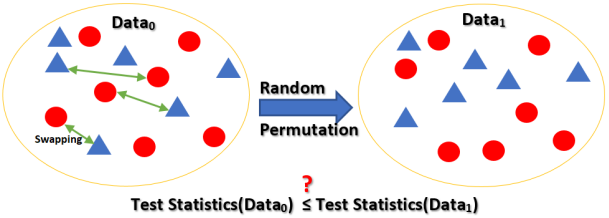


Figure 19: Random permutation test statistics

In this correlation study, the Test Statistics is PIC score for a given radius r. So, in this project in the correlation study, we used random permutation testing to test our hypothesis. Now we are reformulating our main hypothesis:

- **Null Hypothesis(H_0)** The weather events and traffic events have correlation in a given radius r.
- **Alternative Hypothesis(H_1)** The weather events and traffic events are randomly distributed within the given radius r.

We have a set of weather events $\mathbb{W} = \{e_1^w, \cdots, e_n^w\}$, and a set of traffic events $\mathbb{T} = \{e_1^t, \cdots, e_m^t\}$ and $All\_Events = \mathbb{T} \cup \mathbb{W}$. Similar to the process in Figure 19, we did the random permutation process by randomly swapping the event location coordinates in the All_Events set. We set the original data as $Data_0 = (\mathbb{W}, \mathbb{T})$, and we did random permutation process to the original data $Data_0$ p times. We got p data set $\{Data_1, Data_2, \cdots, Data_p\}$, where $Data_i = (\mathbb{W}_i, \mathbb{T}_i)$. For a given radius r, the p_value of the test statistic is the ratio of random permutation data sets whose test statistic scores are not less than the test statistic of the original data $Data_0$ set:

$$p\_value = \sum_{i=1}^{p} \frac{I(PIC(Data_0, r) \leqslant PIC(Data_i, r))}{p}, \tag{4}$$

where r is the test radius, $I(PIC(Data_0, r) \leqslant PIC(Data_i, r))$ is a indicator function, if the inequality is true, then it returns 1, otherwise 0. $PIC(Data_i, r)$ is the PIC score of the $Data_i$ for the given radius r.

We proposed the Algorithm 5, entitled the NETWORK BASED RANDOM PERMUTATION TEST STATISTIC ALGORITHM, to testify our hypothesis. The inputs of the algorithm are a set of weather events $\mathbb{W}$, a set of traffic events $\mathbb{T}$, the testing geographic radius r_g and the time radius r_t. The output is the p_value.

Line 5 calculates the Test Statistic Score of original data $\mathbb{W}, \mathbb{T}$. **Line 7∼16** the outer loop repeats the random permutation test $max\_permutatio\_nnum$ times. **Line 8∼9** in each iteration the location of the events in original set is randomized and randomly permuted new event set is obtained. **Line 10∼11** regroups the new event set into the weather event set and traffic event set (both sets are randomly mixed types of events). **Line 12** calculates the randomly permuted data Test Statistic score. **Line 13∼15** if the randomly permuted data Test Statistic score is equal to or greater than the Test Statistic score then the "total_greater_score_num" is increased by one. **Line 17** calculates the p_value.

Algorithm 5 NETWORK BASED RANDOM PERMUTATION TEST STATISTIC ALGORITHM

1: **Input**: $\mathbb{W} = \{e_1^w, \cdots, e_n^w\}$ //weather event set
2: $\mathbb{T} = \{e_1^t, \cdots, e_m^t\}$ //traffic event set
3: $r_t, r_g, max\_permutation\_num$
4: **Output**: p_value ;
5: $Test\_Statistic\_Score \leftarrow \text{PIC}(\mathbb{W}, \mathbb{T}, r_t, r_g)$
6: $total\_greater\_score\_num \leftarrow 0.0$
7: **for** $i = 1$ to $max\_permutation\_num$ **do**
8: $All\_Events \leftarrow \mathbb{W} \cup \mathbb{T}$;
9: $All\_Events \leftarrow \text{RandomlySwapLocation}(All\_Events)$;
10: $\tilde{\mathbb{W}} \leftarrow All\_Events[1:n]$
11: $\tilde{\mathbb{T}} \leftarrow All\_Events[n+1:n+m]$
12: $Random\_Test\_Statistic\_Score \leftarrow \text{PIC}(\tilde{\mathbb{W}}, \tilde{\mathbb{T}}, r_t, r_g)$
13: **if** $Test\_Statistic\_Score \leqslant Random\_Test\_Statistic\_Score$ **then**
14: $total\_greater\_score\_num \leftarrow total\_greater\_score\_num + 1.0$;
15: **end if**
16: **end for**
17: $p\_value \leftarrow \frac{total\_greater\_score\_num}{max\_permutation\_num}$
18: **return** p_value;

5 Experiments and Discussions

In this section, we will discuss how we designed the real-world data experiments and show the experiment results. We will also analyze the empirical results of the experiments and argue the limitations of our work.

5.1 Experiments

In this section, we will discuss the experiment results of the weather change event detection, the traffic change event detection and the correlation study between the weather event and the traffic event.

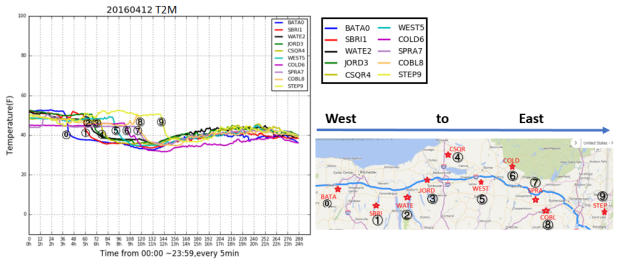


Figure 20: Time-lag effect across the state from West to East [T2M]

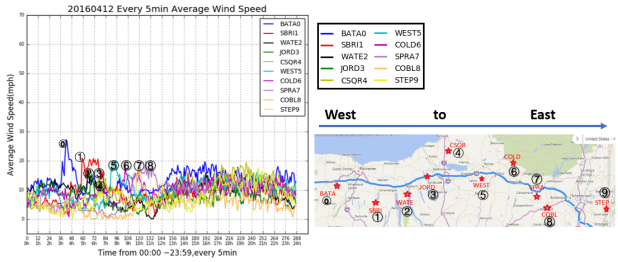


Figure 21: Time-lag effect across the state from West to East [Average Wind Speed]

5.1.1 Time-Lag Effect and Detection

We observed some interesting rapid changes from the weather data plots. Figure 20 shows one example plot for the time-lag effect of T2M change on Apr. 12, 2016. Interestingly Figure 21 shows the corresponding time-lag effect of Average Wind Speed change at the same time on Apr. 12, 2016. The temperature (wind speed) changes across the New York state from West to East. The figure shows that the temperature (wind speed) is dropping down (increasing) one by one on time line from western weather stations to eastern weather stations.

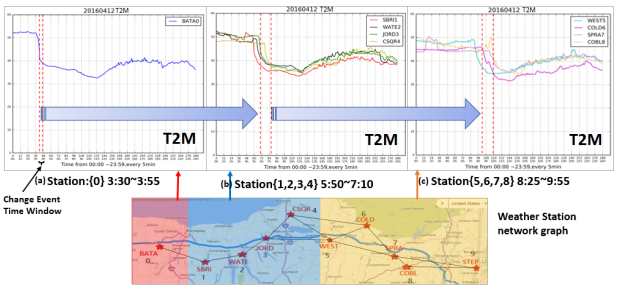


Figure 22: Time-lag effect change event detection

Our proposed method can detect the time-lag effect in rapid weather change detection. If there exists a time-lag effect, then it will be returned as a list of change events. Figure 22 shows the change events corresponding to Figure 20, which shows a time lag effect occurred on Apr. 12, 2016. The (a) subplot in

Figure 22 shows that weather station BATA has a rapid change in temperature when it dropped down very quickly in the time interval 3:30 am $\sim$ 3:55 am. In the plot, we used two red dashed vertical lines to indicate the time window of change event. The (b) subplot shows that two hours later, weather stations SBRI, WATE, JORD, and CSQR also have a rapid weather change in temperature when the temperature dropped down very quickly in the time interval 5:50 am $\sim$ 7:10 am. And the (c) subplot shows that weather stations WEST, COLD, SPRA, and COBL have a rapid weather change in temperature, and the temperature dropped down very quickly in the time interval 8:25 am $\sim$ 9:55 am.

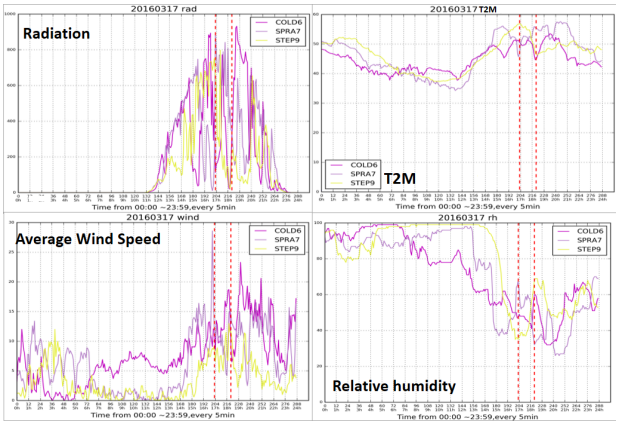


Figure 23: Weather change event detection Case 1

5.1.2 Weather Change Event Detection

In the weather change event detection experiment, we used 8 weather variables data from Mar. 1, 2016 to Dec. 31, 2016 (in total 306 days). As discussed in Section 4.3, in our weather change event detection experiment we applied Algorithm 2 viz. Multi-variable Change Event Detection Algorithm. Algorithm 2 iteratively loads 306 days data of multi-variables and returns a list of change events ranked by their EMS scores. The top ranked change events are most likely to be the true rapid changes.

For each input parameters of Algorithm 2 we tried several different settings. The weather change events do not have ground truth, so we tried different combinations of parameters, $s = \{2, 3\}$, $min\_window = 3$, $max\_window = \{6, 12, 18, 24, 36\}$ and $hist\_base\_length = \{6, 12\}$. After manually checking the true positive change events and false positive change events in the returned top change events, we selected the parameter combination that returned the highest precision score of change events list. So, we set the upper bound s to 2 (our algorithm returns at most $4 * s$ nodes). We set the minimum and maximum time window size min_window and max_window to 3 and 18 respectively, the size of historical base time window $hist\_base\_length$ to 12.

Now we show several case study plots of the multi-variable change event detection experiment:

Case 1 change event: COLD, SPRA, STEP, Radiation, Relative Humidity, T2M, Average Wind Speed, 20160317, 16:55∼18:20.

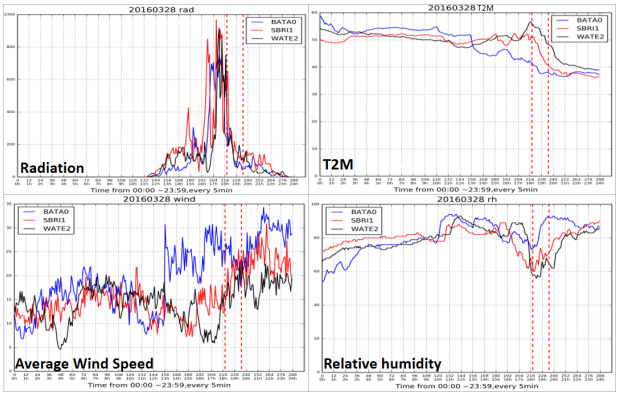


Figure 24: Weather change event detection Case 2

In Case 1 (see Figure 23), there are four plots for four change event related variables viz. the Solar Radiation, T2M, Average Wind Speed and Relative Humidity. The vertical red dash lines indicate the time intervals in which the rapid change occurred. In plots, each different color line represents the values of different stations. Within the time interval 16:55 ∼ 18:20 on Mar. 17, 2016, the Solar Radiation at COLD and SPRA has increased and at STEP weather station it has decreased. The temperature at COLD and STEP has dropped down and at SPRA it has increased. Average Wind Speed at COLD and STEP has increased and at SPRA it has decreased. The Relative Humidity at COLD has increased, at SPRA and STEP it has decreased. We call the value of a variable has decreased or increased within a time window based on the comparison of its time window value and the historical base value.

Case 2 change event: BATA, SBRI, WATE, Radiation, Relative Humidity, T2M, Average Wind Speed, 20160328, 18:10∼19:35.

In Case 2 (see Figure 24), there are four plots for weather change event related variables and at BATA, SBRI, WATE the rapid weather change have occurred within the time interval 18:10 ∼ 19:35 on Mar. 28, 2016. We can see from the plots, the radiation has obviously dropped down very quickly and temperature has also dropped down for all three weather stations. At BATA the Average Wind Speed has decreased and at SBRI, WATE the Average Wind Speed has increased. At all three stations, the Relative humidity has increased.

Case 3 change event: WATE, JORD, WEST, Radiation, Relative Humidity, T2M, 20160529, 18:15∼19:35.

In Case 3 (see Figure 25), there are three plots for weather change event related variables and WATE, JORD, WEST are the weather stations where rapid weather change have occurred within the time interval 18:15 $\sim$ 19:35 on May 29, 2016. From this figure, we can see the obvious weather change events similar to the previous two cases.

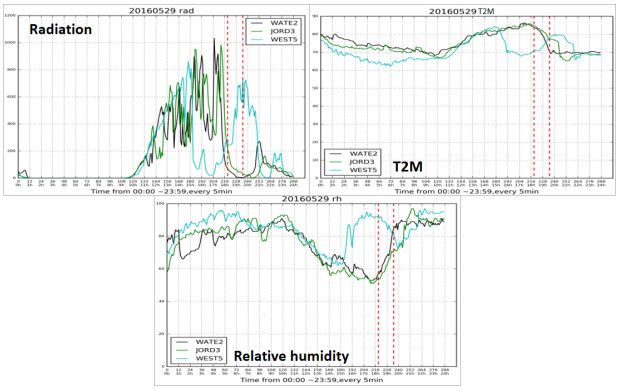


Figure 25: Weather change event detection Case 3

5.1.3 Traffic Change Event Detection

In the traffic change event detection experiment we used the traffic traveling time data for I90 East and I90 West from Mar. 1, 2016 to Dec. 31, 2016 (in total 306 days). As we described in Section 4.2, in traffic change event detection experiment we applied Algorithm 1 viz. Single Variable Change Event Detection Algorithm. In Algorithm 1 we iteratively loaded 306 days data of traveling time, and the algorithm returned a list of change events ranked by their EMS scores. The top ranked change events are most likely to be the true rapid changes.

For the parameters of Algorithm 1 we tried several different settings. The traffic change events also do not have ground truth, so we tried different combinations of parameters, $s = \{2, 3\}$, $min\_window = 3$ $max\_window = \{6, 12, 18, 24, 36\}$ and $hist\_base\_length = \{6, 12\}$. After manually checking for the true positive change events and false positive change events in the returned top change events, we selected the parameter combination that returned the highest precision score of change events list. So, we set the upper bound s to 3 (our algorithm returns at most $4 * s$ nodes). We set the minimum time window min_window and maximum time window max_window sizes to 3 and 18 respectively, the size of historical base time window $hist\_base\_length$ to 12.

From the top change event results we removed the outlier change events related to only one TMC with traveling time greater than 3600 seconds (one hour). Now we show several Case study plots of the single variable change event detection experiment for I90 East and I90 West data.

Case 1 I90 East change event: TMC-46, TMC-47, TMC-48, TMC-49, TMC-50, TMC-51, 5 min Average Traveling Time, 20161214, 19:20$\sim$20:40.

In Case 1 (see Figure 26), at 19:20∼20:40 on Dec. 14, 2016 the traveling times for I90 East TMC-46 ∼ TMC-51 were rapidly increasing.

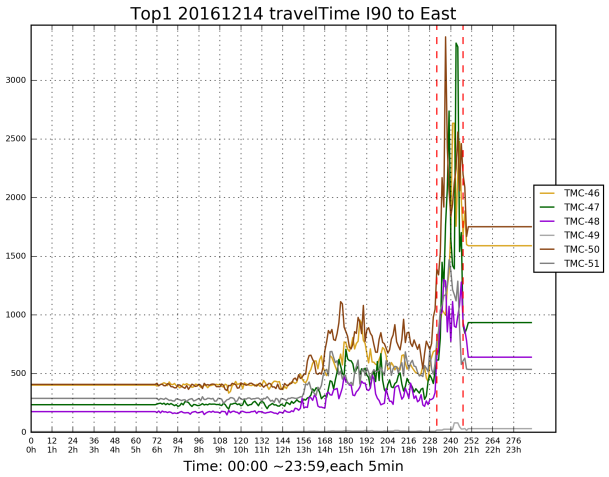


Figure 26: I90 East traffic change event detection experiment Case 1

Case 2 I90 East change event: TMC-40, TMC-41, 5 min Average Traveling Time, 20160916, 9:50∼11:10.

In Case 2 (see Figure 27), at 9:50∼11:10 on Sep. 16, 2016 the traveling times for I90 East TMC-40 and TMC-41 were rapidly increasing.

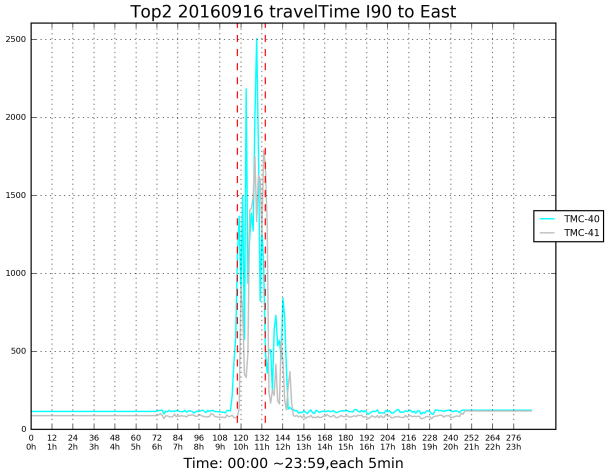


Figure 27: I90 East traffic change event detection experiment Case 2

Case 3 I90 East change event: TMC-18, TMC-19, TMC-20, TMC-21, TMC-22, TMC-23, TMC-24, TMC-25, TMC-26, TMC-27, TMC-28, 5 min Average Traveling Time, 20161215, 15:35∼17:00.

In Case 3 (see Figure 28), at 15:35∼17:00 on Dec. 15, 2016 the traveling times for I90 East TMC-18 ∼ TMC-28 were rapidly increasing.

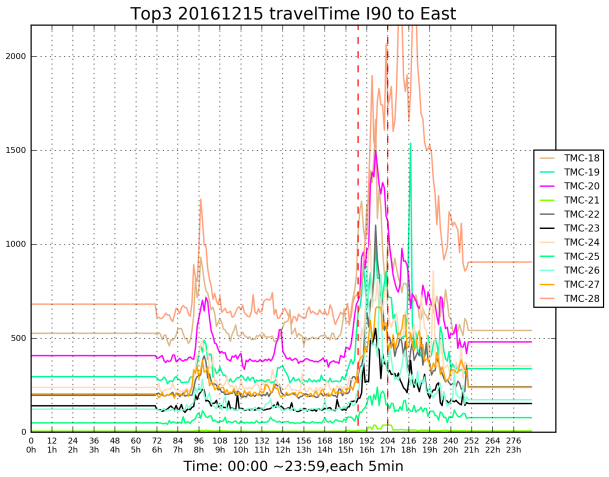


Figure 28: I90 East traffic change event detection experiment Case 3

Case 4 I90 West change event: TMC-45 , TMC-46 , TMC-47 , TMC-48 , TMC-49, 5 min Average Traveling Time, 20161214, 19:25∼20:35.

In Case 4 (see Figure 29), at 19:25∼20:35 on Dec. 14, 2016 the traveling times for I90 West TMC-45 ∼ TMC-49 were rapidly increasing.

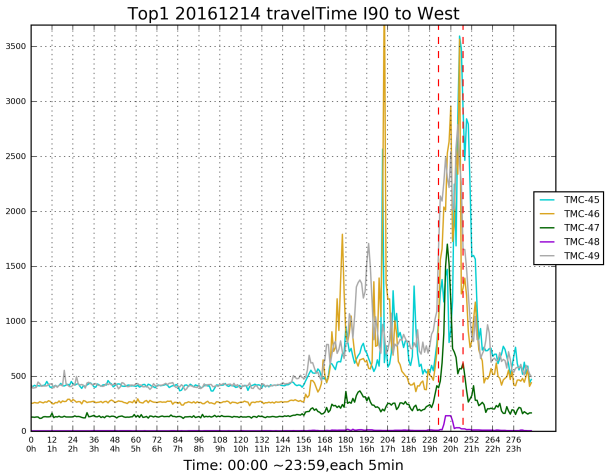


Figure 29: I90 West traffic change event detection experiment Case 4

Case 5 I90 West change event: TMC-17, TMC-18 , TMC-19, TMC-20, TMC-21, TMC-22, TMC-23, TMC-24, TMC-25, TMC-26, TMC-27, 5 min Average Traveling Time, 20161215, 15:35∼17:00.

In Case 5 (see Figure 30), at 15:35∼17:00 on Dec. 15, 2016 the traveling times for I90 West TMC-17 ∼ TMC-27 were rapidly increasing.

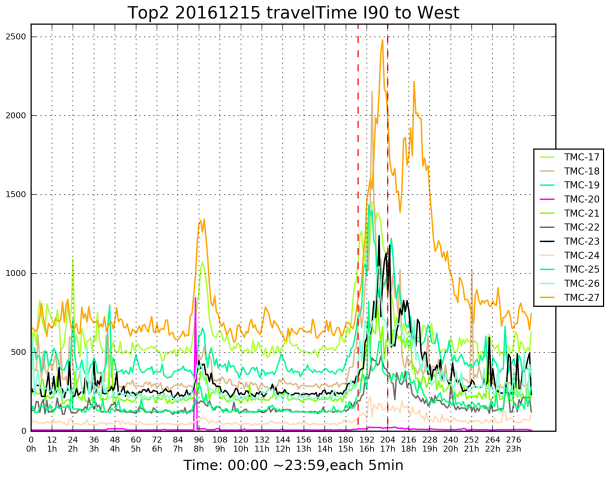


Figure 30: I90 West traffic change event detection experiment Case 5

Case 6 I90 West change event: TMC-45 , TMC-46, 5 min Average Traveling Time, 20161214, 16:45∼17:00.

In Case 6 (see Figure 31), at 16:45∼17:00 on Dec. 14, 2016 the traveling times for I90 West TMC-45 and TMC-46 were rapidly increasing.

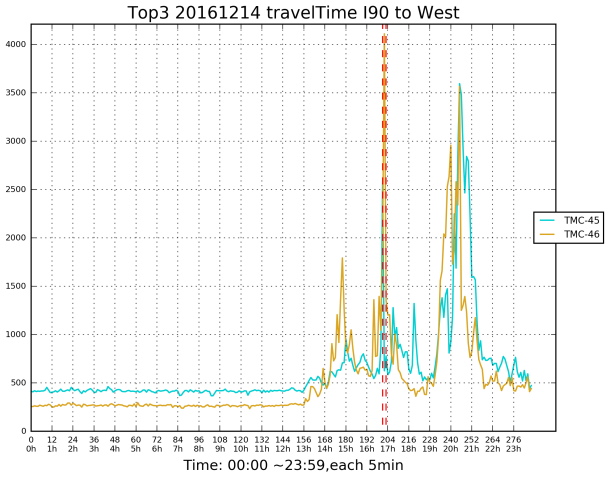


Figure 31: I90 West traffic change event detection experiment Case 6

From the above top ranked change events, we can see that the shapes of traffic change events of neighborhood TMCs are very similar and the rapid changes are quite obvious.

5.1.4 Correlation Study

As we discussed in Section 4.5, we used Top K change event results of both the weather and traffic (includes I90 West and I90 East route results) change event detection experiments to generate the weather events and the traffic events. To determine the values of K for weather change event results and I90 East and I90 West change event results, we calculated the median and mad (mean absolute deviation) of the EMS scores of each group of results separately. We set the value of K to the number of change events that have EMS scores greater than $median + 2 * mad$. We obtained K=103 for the weather change event results, K=33 for the I90 East route change event results and K=29 for the I90 West route change event results. Table 3 shows Top 103 weather change events, Table 4 shows the Top 33 I90 West traffic change events and Table 5 shows the Top 29 I90 East traffic change events.

Table 3 Top Ranked Weather Change Event List

Rank	Weather Variables	Weather Stations	Date	Time Slot
1	T2M, T9M, Average Wind Speed, Max. Wind Speed, Relative Humidity	SBRI, WATE, JORD, CSQR	20160908	172∼186
2	T2M, T9M, Average Wind Speed, Max. Wind Speed, Relative Humidity	CSQR, COLD, SPRA, STEP	20160317	171∼188
3	T2M, T9M, Average Wind Speed, Max. Wind Speed, Relative Humidity	SBRI, WATE, JORD, CSQR	20161119	164∼178
4	T2M, T9M, Average Wind Speed, Max. Wind Speed, Relative Humidity	JORD, CSQR, WEST, SPRA	20160402	181∼197
...	...	...	...	...
102	T2M, T9M, Relative Humidity	SBRI, WATE, JORD, WEST	20160906	225∼240
103	T2M, T9M, Relative Humidity	WATE, JORD, CSQR, COLD	20160526	142∼159

Table 4 I90 West Top Ranked Traffic Change Event List

Rank	TMC List	Date	Time Slot Interval
1	TMC-45, TMC-46, TMC-47, TMC-48, TMC-49	20161214	233∼247
2	TMC-17, TMC-18, TMC-19, TMC-20, TMC-21, TMC-22, TMC-23, TMC-24, TMC-25, TMC-26, TMC-27	20161215	187∼204
3	TMC-45, TMC-46	20161214	201∼203
...	...	...	...
28	TMC-9, TMC-10, TMC-11	20160508	139∼156
29	TMC-8, TMC-9	20160917	113∼130

After selecting the Top K weather weather change events and Top K traffic change events, we used them to generate the weather events and traffic events by applying the method introduced in Section 4.5.1. We obtained 5755 weather events and 3657 traffic events (total events of I90 East and I90 West route). Each event has Event Type, Latitude, Longitude, Time and Station or TMC ID. We re-indexed the Weather Station IDs from 100 to 109, I90 East TMC IDs from 200 to 253 and I90 West TMC IDs from 300 to 351. For example, consider an event (Event Type :0, Location: (42.75, -77.36), Date and Time slot:(20160908,172), ID:101). This is a weather event that occurred at weather

Table 5 I90 East Top Ranked Traffic Change Event List

Rank	TMC List	Date	Time Slot Interval
1	TMC-46, TMC-47, TMC-48, TMC-49, TMC-50, TMC-51	20161214	232 ∼ 247
2	TMC-40, TMC-41	20160916	118∼ 134
3	TMC-18, TMC-19, TMC-20, TMC-21, TMC-22, TMC-23, TMC-24, TMC-25, TMC-26, TMC-27, TMC-28	20161215	187∼204
...	...	...	...
31	TMC-8, TMC-9, TMC-10	20160514	232∼249
32	TMC-27, TMC-28, TMC-29, TMC-30, TMC-31, TMC-32, TMC-33, TMC-34, TMC-35, TMC-36, TMC-37	20160403	233∼ 250
33	TMC-1, TMC-2	20160920	86∼101

station with ID 101 and the geographic location coordinate of the station is (42.75,-77.36), the event time consists of date Sep. 8, 2016 and time slot 172 (14:20 PM).

In the correlation study experiment, we applied the Algorithm 5 (Network based Random Permutation Test Statistic Algorithm). Algorithm 5 takes the weather events, traffic events and maximum permutation number parameter $max\_permutation\_num$=500 as input. For the geographic radius r_g we tested $\{5, 10, 15, \cdots, 45\}$ (the unit is miles) and for the time radius r_t we tested $\{1, 2, 3, 4, 5\}$ time slots (or $\{5, 10, 15, 20, 25\}$ minutes). Based on our assumption and PIC spatial pairwise correlation statistic, if two events have correlation on the given fixed radius (or distance), then we expect the p-values to be minimum at the given geographic radius and time radius.

Figure 32 shows the results of our correlation study experiments. When the geographic radius is less than or equal to 10 miles, the random permutation test returned the minimum p-values. The weather event and traffic event have a correlation in given radius less than or equal to 10 miles.

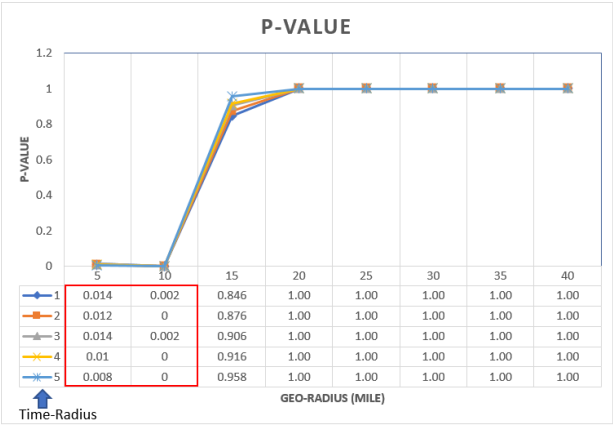


Figure 32: Correlation study experiment result

5.2 Discussions and Limitations

In this project, our main goal was to detect change events from a weather station network and the TMC network, further, we used detected change events to examine the correlation between rapid weather change events and rapid traffic change events. To the best of our knowledge, this is the first work to study the relations between rapid weather change and rapid traffic change. The previously existing methods input the weather or traffic data into the model without considering the spatio-temporal rapid changes of the data. We modeled the weather stations and TMCs into a network graph and found a new approach to view the raw data. We also discovered more interesting rapid changes from the raw data (see Section 5.1.1).

We conducted comprehensive experiments and determined the best parameter settings for our proposed change event detection algorithms. In our change event detection experiment, our two main proposed algorithms Algorithm 1 viz. single variable change event detection algorithm and Algorithm 2 viz. multivariable change event detection algorithm performed very well. In Section 5.1.2 and Section 5.1.3 we showed the top k results of our change event experiments and manually justified each of the top K change events.

In the correlation study, the input data of the experiment comes from previous change event detection experiments. We have already discussed the quality of change events. We manually checked and ensured that all of the change events are obvious change events. We tried different parameter settings when we run Algorithm 5 to obtain expected results. When the testing geographic radius is less than or equal to 10 miles, the p-values are lower than 0.03. What would be the impact of the final results of change event detection experiments on the correlation study experiment? We used Mesonet weather data and NPMRDS Traffic data, however, both datasets have missing values. To reduce the effect of missing data on the result of change event detection, we set the previous time slot value (for which the data are available) as the value for the time slot with missing data, so the new extrapolated data values and the historical base values have the same or close value. This data interpolation process diminishes the effect of the missing data. Also, we only have data from Mar. 2016 to Dec. 2016 for both weather and traffic datasets. We believe that access to a larger dataset e.g. data for two or more years of time, having more variables e.g. average vehicle speed, I90 accidents etc. in case of traffic data and precipitation, snow coverage, visibility distance etc. in case of weather data will improve the accuracy of our experiments. Also, getting data from weather stations that are close to I90 routes will provide us with more reliable input and further improve the accuracy of our technique. Further, getting the ground truth for weather and traffic change events or finding more suitable distributions in our assumptions in the hypothesis tests will lead to more accurate and convincing results.

6 Conclusion

The goal of this project is to study the correlation between rapid weather change events and rapid traffic change events by using graph based change event detection algorithms. We also designed a new network based random permutation method to study the spatio-temporal correlations of the weather events and the traffic events. We proposed three novel algorithms: Algorithm 1 is a single variable change event detection algorithm; Algorithm 2 is a multi-variable change event detection algorithm and Algorithm 5 is a network based random permutation test statistic algorithm.

We used the weather data and traffic data in our network based spatio-temporal model and found a new angle to analyze the weather data and traffic data. Our case study experiment results showed that our proposed algorithms performed very well. The proposed research showed the feasibility of using weather data and traffic data to detect the traffic condition change during adverse weather conditions in real-time or further, predict the traffic condition change during adverse weather conditions. In Experiments Section 5, we can see that the proposed change event detection algorithms, Algorithm 1 and 2, are very flexible and we can use the same change event detection methods to 8 different weather variables data and also the traveling time data. The experiment results show high quality change events. Also, our correlation study experiment shows the reasonable results. When the testing radiuses are less than or equal to 10 miles, our random permutation test statistic return the minimum p-values, which indicates the weather events and traffic events have a high correlation when their distances are less than or equal to 10 miles..

The methodology used in this research promises a way forward for the development of techniques to be applied to real-time data feeds. By applying our technique to real-time 5-minute weather observations and real-time 5-minute traffic data, we anticipate the ability to alert traffic operations staff, decision makers, and the public, to upcoming conditions. Further research will examine the relationship between conditions identified using our rapid-change algorithm with weather and traffic data, and recorded incidents, to inform safety researchers and transportation planner.

References

[1] P. A. Pisano, L. C. Goodwin, and M. A. Rossetti, "U.S. highway crashes in adverse road weather conditions," in *24th Conference on International Interactive Information and Processing Systems for Meteorology, Oceanography and Hydrology, New Orleans, LA*, 2008.

[2] "How do weather events impact roads? retrieved from 25 usdot fhwa road weather management program:," 2014.

[3] "U.S. department of transportation, federal highway administration," https://ops.fhwa.dot.gov/weather/q1_roadimpact.htm.

[4] S. Liu, M. Yamada, N. Collier, and M. Sugiyama, "Change-point detection in time-series data by relative density-ratio estimation," *Neural Networks*, vol. 43, pp. 72–83, 2013.

[5] J. Bai, "Estimation of a change point in multiple regression models," *Review of Economics and Statistics*, vol. 79, no. 4, pp. 551–563, 1997.

[6] Y. Kawahara, T. Yairi, and K. Machida, "Change-point detection in time-series data based on subspace identification," in *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on.* IEEE, 2007, pp. 559–564.

[7] D. B. Neill, E. McFowland, and H. Zheng, "Fast subset scan for multivariate event detection," *Statistics in medicine*, vol. 32, no. 13, pp. 2185–2208, 2013.

[8] R. Jiang, H. Fei, and J. Huan, "A family of joint sparse pca algorithms for anomaly localization in network data streams," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 25, no. 11, pp. 2421–2433, 2013.

[9] M.-A. Tessier, C. Morency, and N. Saunier, "Impact of weather conditions on traffic: Case study of montreals," in *TRB*, 2015.

[10] R. Lamm, E. M. Choueiri, and T. Mailaender, "Comparison of operating speeds on dry 10 and wet pavements of two-lane rural highways," in *Transportation Research Record 1280*, 1990.

[11] S. Datla and S. Sharma, "Impact of cold and snow on temporal and spatial variations of highway traffic volumes," *Journal of Transport Geography*, vol. 16, no. 5, pp. 358–372, 2008.

[12] L. Zhao and S. I.-J. Chien, "Analysis of weather impact on travel speed and travel time reliability," in *CICTP 2012@ sMultimodal Transportation SystemsConvenient, Safe, Cost-Effective, Efficient.* ASCE, 2012, pp. 1145–1155.

[13] T. Kwon, L. Fu, and C. Jiang, "Effect of winter weather and road surface conditions on macroscopic traffic parameters," *Transportation Research Record: Journal of the Transportation Research Board*, no. 2329, pp. 54–62, 2013.

[14] W. Van Stralen, "The influence of adverse weather conditions on the probability of congestion on dutch highways," Ph.D. dissertation, TU Delft, Delft University of Technology, 2013.

[15] M. Elhenawy, H. Chen, and H. A. Rakha, "Traffic congestion identification considering weather and visibility conditions using mixture linear regression," in *Transportation Research Board 94th Annual Meeting*, no. 15-3323, 2015.

[16] M. Martchouk, F. L. Mannering, and C. NEXTRANS, "Analysis of travel time reliability on indiana interstates," NEXTRANS, Tech. Rep., 2009.

[17] E. I. Vlahogianni, M. G. Karlaftis, and J. C. Golias, "Short-term traffic forecasting: Where we are and where were going," *Transportation Research Part C: Emerging Technologies*, vol. 43, pp. 3–19, 2014.

[18] Y.-J. Wu, F. Chen, C.-T. Lu, and S. Yang, "Urban traffic flow prediction using a spatio-temporal random effects model," *Journal of Intelligent Transportation Systems*, pp. 1–12, 2015.

[19] B. L. Smith and M. J. Demetsky, "Traffic flow forecasting: comparison of modeling approaches," *Journal of transportation engineering*, vol. 123, no. 4, pp. 261–266, 1997.

[20] B. L. Smith, B. M. Williams, and R. K. Oswald, "Comparison of parametric and nonparametric models for traffic flow forecasting," *Transportation Research Part C: Emerging Technologies*, vol. 10, no. 4, pp. 303–321, 2002.

[21] A. Stathopoulos and M. G. Karlaftis, "A multivariate state space approach for urban traffic flow modeling and prediction," *Transportation Research Part C: Emerging Technologies*, vol. 11, no. 2, pp. 121–135, 2003.

[22] X. Min, J. Hu, and Z. Zhang, "Urban traffic network modeling and short-term traffic flow forecasting based on gstarima model," in *Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference on*. IEEE, 2010, pp. 1535–1540.

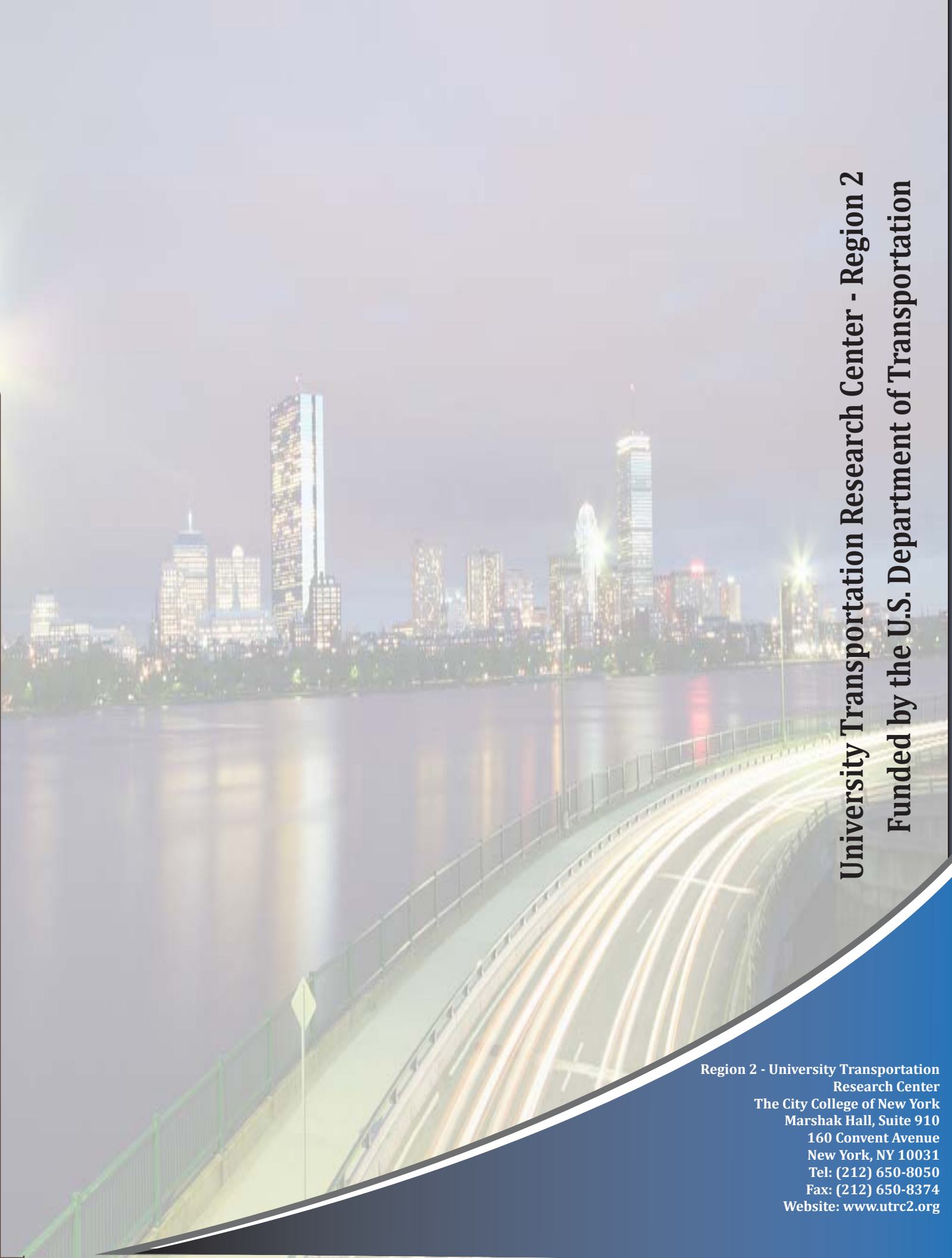
[23] Z. Ghahramani and G. E. Hinton, "Variational learning for switching state-space models," *Neural computation*, vol. 12, no. 4, pp. 831–864, 2000.

[24] "Mesonet website," http://www.nysmesonet.org.

[25] F. Chen and B. Zhou, "A generalized matching pursuit approach for graph structured sparsity," *The 25th International Joint Conference on Artificial Intelligence (IJCAI'16)*, 2016.

[26] ——, "Graph-structured sparse optimization for connected subgraph detection," *The 22st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16) (In Review)*, 2016.

[27] S. Barua and J. Sander, "Mining statistically sound co-location patterns at multiple distances," in *Proceedings of the 26th International Conference on Scientific and Statistical Database Management*. ACM, 2014, p. 7.



University Transportation Research Center - Region 2

Funded by the U.S. Department of Transportation

Region 2 - University Transportation
Research Center
The City College of New York
Marshak Hall, Suite 910
160 Convent Avenue
New York, NY 10031
Tel: (212) 650-8050
Fax: (212) 650-8374
Website: www.utrc2.org