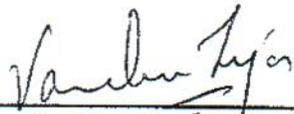


APPROVAL SHEET

Title of Thesis: Discovery of Multi-Domain Anomalous Temporal Association

Name of Candidate: Suraksha Shukla
Master of Science, Information System 2017

Thesis and Abstract Approved:



Dr. Vandana Janeja
Associate Professor
Information System

Date Approved:

5/9/17

ABSTRACT

Title of Document: DISCOVERY OF MULTI-DOMAIN
ANOMALOUS TEMPORAL
ASSOCIATIONS.

Suraksha Shukla
Master of Science, Information Systems.

Directed By: Associate Professor, Dr. Vandana Janeja,
Information Systems

Temporal data can capture the behavior of phenomena such as accidents along a highway, weather trend such as precipitation or snow totals in a region over time. Traditional temporal data mining has looked at patterns, such as anomalies, in each temporal data stream. However, to study real world phenomena and inter relationships between them, in this thesis, we propose a novel approach to discover the temporal relations between multiple distinct domains represented by multiple distinct temporal data collected at a location. Our goal is to discover the relationship between distinct domains using interesting temporal events in them. These interesting temporal events are mined using traditional temporal anomaly detection methods. Relations between two application domains are not always simple since there can be some time-delay in these relationships. So, focusing on relations found using intersecting time events alone is not sufficient. Hence, we employ the concept of not only direct overlap but also

proximity between temporal events across domains to find the direct and time-delayed relationships. We have achieved an optimistic result after our experiment on MATCH, NJDOT and weather data.

DISCOVERY OF MULTI-DOMAIN ANOMALOUS TEMPORAL
ASSOCIATIONS.

By

Suraksha Shukla

Thesis submitted to the Faculty of the Graduate School of the
University of Maryland, Baltimore County, in partial fulfillment
of the requirements for the degree of
Master of Science,
Information System
2017

© Copyright by
Suraksha Shukla
2017

Dedication

I dedicate this thesis to my beloved family.

Table of Contents

Dedication	ii
Table of Contents	iii
List of Tables	v
List of Figures	vi
Chapter 1: Introduction	1
1.1 Overview	1
1.1.1 Multi-domain Temporal Anomaly Detection	1
1.1.2 Challenges	2
1.2 Motivation	2
1.3 Thesis Contribution:	5
Chapter 2: Related Work	6
2.1 Temporal Interpolation:	6
2.2 Anomaly Detection	8
2.3 Temporal Association Rule Mining:	8
Chapter 3: Methodology	12
3.1 Data Preprocessing:	14
3.1.1 Temporal Interpolation	14
3.2 Discovery of anomalies	16
3.3 Association of overlapped anomalous windows	19
3.4 Time-delayed relation	25
3.5 Accuracy Measures	28
Chapter 4: Experiment and result	30
4.1 Software and packages specification	30
4.2 Datasets	30
4.2.1 MATCH Health Ranking data	31
4.2.2 NJDOT data	33
4.2.3 Synthetic data	36
4.2.4 Weather data:	37
4.3 Experiment and results	37
4.3.1 Data Preprocessing	38
4.3.1.1 Temporal Interpolation	38
4.3.2 Single domain anomalies	39
4.3.2.1 Discovered anomalies in MATCH data	39
4.3.2.2 Discovered anomalies in NJDOT data	40
4.3.2.3 Discovered anomalies in synthetic data	42
4.3.2.4 Discovered anomalies in weather data:	44
4.3.3 Temporal Anomalies Association	46
4.3.3.1 Association in MATCH data	47
4.3.3.2 Association in NJDOT data	47
4.3.3.3 Association in synthetic data	49
4.3.4 Time-delayed Association	51

4.3.5 Performance Comparison and Accuracy Evaluation	53
4.3.5.1 Contingency Table	53
4.3.5.2 Correlation	57
4.3.5.3 PAA.....	59
4.3.5.4 Comparison with larger dataset	62
Chapter 5: Conclusion and Future works.....	65
Bibliography	66

List of Tables

Table 1: Anomalous time events in Accident and Light Condition domains	4
Table 2: Summary of Terminologies used.....	14
Table 3: Precision, Recall and Accuracy formula	29
Table 4: Measures used to rank county’s health condition.....	33
Table 5: List of files used for MATCH data.....	33
Table 6: Categorized variables in NJDOT data	34
Table 7: Light condition codes	35
Table 8: Surface Condition Code.....	35
Table 9: List of files used for NJDOT data.....	36
Table 10: List of files used for synthetic data.....	36
Table 11: List of files used for weather data.....	37
Table 12: Details of Datasets Used.....	38
Table 13: Anomalies in domains of MATCH data.....	40
Table 14: Anomalies in domains of NJDOT data.....	41
Table 15: Inserted anomalous windows in synthetic data	43
Table 16: Anomalies in synthetic data.....	43
Table 17: Anomalies in weather data.....	45
Table 18: Support, Confidence and Lift in NJDOT data	48
Table 19: Support, Confidence and Lift in synthetic data	49
Table 20: Delayed relation found in weather data	53
Table 21: Anomalies in real-world and synthetic data of Total Injured.....	54
Table 22: Contingency table for Total Injured	54
Table 23: Anomalies in real-world and synthetic data of Light Condition	55
Table 24: Contingency table for Light Condition.....	55
Table 25: Anomalies in real-world and synthetic data of Surface Condition.....	55
Table 26: Contingency table for Surface Condition	55
Table 27: Precision, Recall and Accuracy	56
Table 28: Correlation for all domain pairs in each bin (NJDOT).....	58
Table 29: Correlation for all domain pairs in each bin (Synthetic data).....	59
Table 30: Anomalies discovered in one year data (NJDOT).....	63

List of Figures

Figure 1: Anomalies in accident and light condition domain	3
Figure 2: State of the art interpolation techniques	8
Figure 3: State of the art temporal mining techniques	11
Figure 4: Multi-domain Temporal Association Process	13
Figure 5: Anomaly detection process	17
Figure 6: Anomalies in time series domain D1	18
Figure 7: Anomalous time window overlap.....	20
Figure 8: Proximity of $2t$	21
Figure 9: Multi-domain anomaly association framework.....	22
Figure 10: Sample transaction file used for association	23
Figure 11: Time-delayed multi-domain anomaly association	26
Figure 12: Contingency Table	29
Figure 13: Community health ranking approach	32
Figure 14: Interpolation techniques accuracy comparison plot	39
Figure 15: Anomalies in MATCH data	40
Figure 16: Anomalies in NJDOT data for all domains	42
Figure 17: Anomalies in synthetic data for all domains	44
Figure 18: Anomalies plot for weather data	46
Figure 19: Relations with time information.....	48
Figure 20: Relations with time information.....	50
Figure 21: Correlogram between Temperature and Humidity in Bin 2.....	52
Figure 22: Precision, Recall and Accuracy plot.....	56
Figure 23: PAA plot for all domains (NJDOT)	60
Figure 24: PAA plots for all domains (Synthetic data).....	61
Figure 25: PAA plot for Solar Radiation, Temperature, and Humidity (Weather)	62
Figure 26: Anomalies in whole year data (NJDOT).....	63

Chapter 1: Introduction

1.1 Overview

1.1.1 Multi-domain Temporal Anomaly Detection

Time series data mining is a widely-studied area of research. Anomaly detection in time series is one of the fundamental subjects in data mining that has attracted enormous attention of many researchers over the past couple of years. Traditional temporal anomaly detection techniques identify the pattern of a single time series data. However, detected unusual behavior in one variable can have impacts on other variables as well (Janeja, 2013) and analyzing a time series data as an independent feature cannot identify the complex nature of real world problems such as accidents along a highway, traffics in network nodes, poor health condition in a county and so on. These problems are intertwined across other multiple disparate domains and to discover the complete pattern, relationships between the intertwined domains must be discovered as well. Analyzing multiple distinct domains also possesses a unique challenge due to the heterogeneity of data. Hence, a sophisticated technique that identifies and quantifies relations between multiple inter-related domains is needed.

In this thesis, our goal is to analyze temporal anomalies for multi-domain datasets to discover relations among anomalies of these distinct domains. We have proposed a novel algorithm to discover the temporal relations between multiple distinct domains using anomalous time windows. To further complement our approach, we also employ

the concept of overlap and proximity to identify delayed relationships between multiple distinct domains.

1.1.2 Challenges

Multi-domain temporal data mining is the technique to analyze data of complex disparate domains. Finding data of such kind is a challenge in itself. Once found, it requires rigorous data cleaning and transformations. As we are using multiple domains, we also have to deal with the complication of data heterogeneity of all domains. We need to handle all these aspects of data quite carefully as it can have a significant impact on the performance of our association algorithm. Discovery of association patterns of multiple domains needs a framework that does a comprehensive analysis to properly capture all possible cases of temporal relations as simultaneous impacts and delayed impacts.

1.2 Motivation

Multi-domain anomaly detection allows us to unveil complex patterns which are not quite possible to achieve using traditional data mining techniques. This technique is highly useful for time-series data of various application domains where the context of time aspect is very vital for analysis. For example, it allows to uncover temporal patterns such as (a) high traffic on a computer node during office hours is normal but the same level of traffic at 3 am is an anomaly, (b) 45°F temperature in Maryland during the month of January is expected but not in month of July, (c) high number of accidents in a road section during snow season, (d) low quality of health during the time of recession, and so on. Also, another very crucial aspect to heed is, anomalies in these

domains are more or less impacted by some other application domains. Considering the impact and quantifying the level of impact of other domains leads to a more accurate analysis and result while also revealing actual challenges. Multi-domain anomalies generally herald a potential problem which later can be researched for more ground truth and taken care of.

Hence, our goal is to discover temporal anomaly for multi-domain data as we believe that we get some intriguing time associated results. Then we analyze those discovered anomalies further for two conditions, first we see if there is an overlap between those anomalous time sequence, and second, if there is no overlap then we check if those anomalies of different domains are within the specified proximity, if they are then we measure their time delayed correlation. For both cases, for further analysis, we use association rule mining technique to get the relation between those domains.

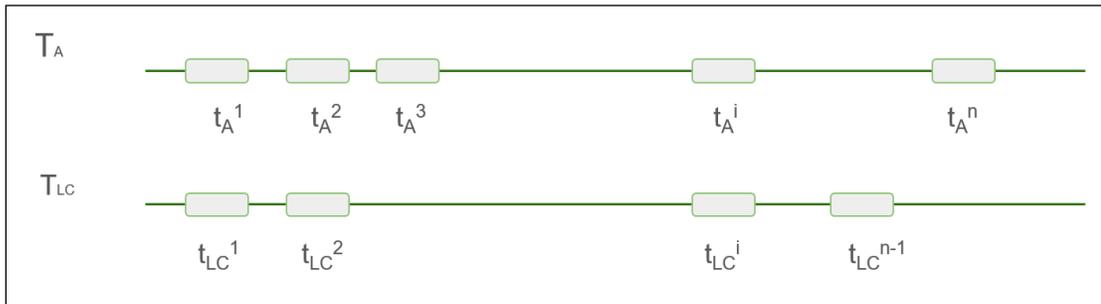


Figure 1: Anomalies in accident and light condition domain

As a motivational scenario let's consider the transportation domain: a high number of accidents in a road section during the month of January may indicate some defect on road condition. For example, a road being covered by snow and drivers having a hard time staying in their lane due to hidden road markers and road being too slippery. However, without considering other domains like weather or light condition that might

have impacted the accidents, uncovering the relevant patterns is not feasible. Figure 1 shows the anomalous data points in accident domain which is represented by $\{t_A^1, t_A^2, t_A^3, \dots, t_A^i, \dots, t_A^n\}$ and in light condition domain which is represented by $\{t_{LC}^1, t_{LC}^2, \dots, t_{LC}^i, \dots, t_{LC}^{n-1}\}$. Here we can see that there are overlaps between unusual events in these domains as shown in table 1. These overlaps are potentially a harbinger of some relation between these two domains. To be more certain about the impact of the light condition on the number of accidents, we use our novel algorithm to discover and quantify the relation between multiple distinct domains.

Domains	Anomalous time events	Overlapping time events
Number of accidents	$t_A^1, t_A^2, t_A^3, \dots, t_A^i, \dots, t_A^n$	t_A^1, t_A^2, t_A^i
Light Condition	$t_{LC}^1, t_{LC}^2, \dots, t_{LC}^i, \dots, t_{LC}^{n-1}$	$t_{LC}^1, t_{LC}^2, t_{LC}^i$

Table 1: Anomalous time events in Accident and Light Condition domains

Discovery of these multi-domain relations can give Department of Transportation a good place to start to further look into the impact of light conditions on accidents. Using the temporal anomaly detection technique, anomalies in both domains, accidents, and light condition, can be discovered, which would be the time sub-sequence of unusually low or high number of accidents for the accident domain and time sub-sequence of unusual light condition for the light condition domain. Then, we look at how they are related, that is if there is an overlap or time delayed correlation. If there is a time delayed correlation, then we keep time series of one domain constant and shift the other domain by the time-delayed coefficient by a certain width. We then use our algorithm

for association rule mining to identify the relation and quantify them using various accuracy measures.

1.3 Thesis Contribution:

In this thesis, we propose an algorithm to discover temporal relationships among multiple associated domains using the anomalous time-sequence in those domains for overlapped and time-delayed relationships. Our major contributions are as follows:

- We propose a novel algorithm for discovering relations between multiple distinct domains using the temporal association on anomalous windows from those domains.
- We employ the concept of overlap and proximity for discovering temporal associations.
- To discover the delayed relation between multiple distinct domains, we use delayed correlation and we shift one domain by time-delayed coefficient δ , to study the relations we use our novel algorithm.
- For quantifying the association, we use confidence, support, and lift. To quantify the performance of anomaly detection technique we utilize accuracy, precision, and recall on synthetic data.
- We also use correlation technique as an added layer of validation for domains with overlapping anomalous window pairs.

Chapter 2: Related Work

We reviewed several research papers on temporal interpolation, anomaly detection, and temporal association rule mining to get some insight on state of the art techniques. In this chapter, we discuss the techniques and their application presented in papers we reviewed. We have categorized the review of papers in three section – temporal interpolation, anomaly detection, and temporal association rule mining.

2.1 Temporal Interpolation:

The missing data values problem is very common in every kind of data, but for a time series data, another issue of time resolution also exists if data from multiple domains recorded in the different time interval are being used. Missing data problem can be resolved by deleting records with missing values but for data in the different resolution, that is not an answer as one cannot afford to delete the data if an analysis is to be done in finer granularity. However, the time series with higher granularity can be interpolated to get intermediate values. The linear interpolation method is the most common and widely used technique for interpolation but has the drawback of highly smoothing out values (Gorman, 2009). For temporal analysis, often sharp variations due to multiple associated features are expected. Moreover, complex application domains with missing data in more than one variable pose a unique challenge (Buuren, 2011). So, an advanced and robust interpolation technique is required that takes into account the correlation between multiple variables. We reviewed few such advanced interpolation techniques that were used for temporal analysis. Lin et al. (2003) proposed a very simple yet sophisticated method, Piecewise Aggregate Approximation

(PAA) which divides a time series data into segments and records a mean value of the data points that fall within the segment, and that mean value then can be used as fill the missing value in a data. Gorman (2009) compared the performance of three interpolation methods, Linear interpolation, Fast Fourier Transform (FFT), and Empirical Orthogonal function (EOF) for wind field data. Linear interpolation and FFT performance were found satisfactory for wind fields that were moving with the constant translation velocity. However, for more complex situation containing multiple evolving weather system, only EOF method was capable of capturing modes of variation, hence performed substantially well. Zavala-Hidalgo et al. (2003) used Empirical Orthogonal Function (EOF) method for vector wind fields to generate the values at intermediate times for a coarsely sampled temporal resolution (Zavala-Hidalgo, 2003). Mice is an R library that imputes missing values for a multivariate data using chained equation (Buuren, 2011).

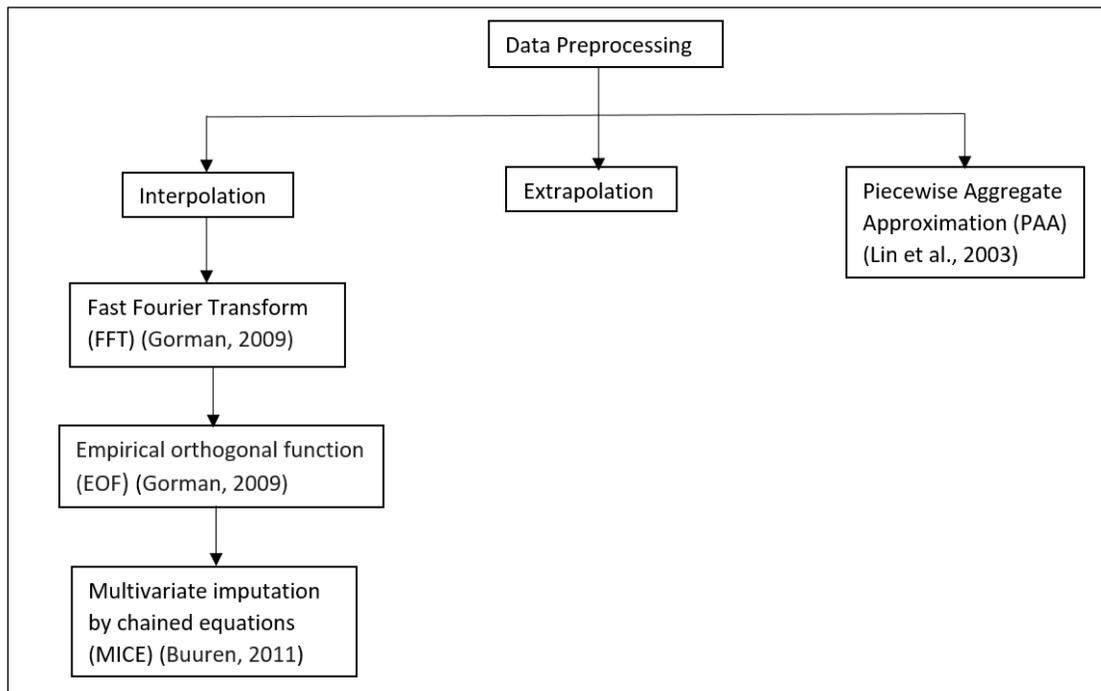


Figure 2: State of the art interpolation techniques

2.2 Anomaly Detection

Legendary discoveries such as Quasars, radio pulsars, and cosmic gamma-ray bursts are results of scientists encountering aberrant phenomena while examining data closely. But with the amount of data that we have these days, analyzing these data for such discovery can be too tedious and almost impossible (Rebbapragada et al., 2009). Therefore, various data mining techniques are used to ease and speed up the process. Anomaly detection is the process of discovering deviations or unusual behavior from normalcy (Chan et al., 2005). An anomaly in a temporal data is a novel or unseen behavior or sequence of behaviors that is different from the rest of the time-series data which are considered to be normal (Teng, 2010). Anomaly detection for a time series data for a single domain has been addressed by many researchers (Teng, 2010, Golmohammadi, 2015). Approaches that has been used for anomaly detection in past are window based, proximity based, prediction based, Hidden Markov Model based, segmentation based and so on (Golmohammadi, 2015). Chan et al.2005 used the Greedy – Split algorithm for anomaly detection in a single time series. Data transformation before detecting the anomaly and the anomaly detection technique are two dimensions of anomaly detection in time series (Golmohammadi, 2015).

2.3 Temporal Association Rule Mining:

Discovering hidden relations between sequences and subsequences of events is the goal of temporal data mining (Antunes et al., 2001). These relations are discovered using association rule mining. The concept of association rule mining was first introduced by

Agrawal et al. (1993) for market basket data. Association rule mining gives relations for items occurring together frequently. Apriori is the most common algorithm used to discover interesting relations. Roddick et al. (2001) used the Apriori-like method, causal rule on temporal data to discover rules comprising time information (Roddick et al., 2001). As compared to conventional association rule mining, temporal association rule adds time information which might be a time point or time range (Liang et al., 2005). Time information is of paramount importance while mining associations for time-varying data sets (Nair et al., 2015). A typical rule $A \Rightarrow B$ (if A occurs then B occurs) changes to $A \Rightarrow Y^T$ (if A occurs then B occurs during the time interval T) after adding time constraint. This allows us to look into chunks of timeframes separately and different rules can be found for different timeframes (Antunes et al., 2001). Adding temporal information leads to some different approaches. Therefore, support and confidence also need to be modified. Antunes et al. (2001) suggested support as “the fraction of entities, which had consumed the itemsets in any of their possible transactions”. Harms (2005) used Episodal association mining for discovering periodic occurrence of interesting events. Harms et al. (2001) used Gen-FCE method to discover the frequent and closed set of episodes (FCE), where an episode is a combination of events with a given order (Harms, 2005). Ramaswamy et al. (1998) used Calendric Association Rule, which is an optimization on “cyclic association rule” to capture real-life complicated temporal patterns. Authors used calendar algebra to capture such real-life event patterns in the form of algebraic expressions. They defined the support for an itemset X in $T[j]$ as the fraction of transactions comprising of the itemset X where $T[j]$ is the set of transactions during the time unit t_j . Confidence for the rule $X \Rightarrow Y$ for

transaction $T[j]$ was defined as the fraction of transactions in $T[j]$ containing X that also contains Y . Zhou et al. (2008) used Genetic Network Programming (GNP) for time series association rule mining on traffic data and has used chi-square and support to measure the importance of association rules. Nair et al. (2015) also used support in their approach where they used Symbolic Aggregate approXimation (SAX)–Apriori based stock trading recommender system to mine temporal association rules for stock price data. However, (Antunes et al., 2001, Nair et al., 2015, Zhou et al., 2008) have not addressed adapting confidence in temporal mining. Liang et al. (2005) used T-Apriori algorithm, which is a modification of the Apriori algorithm, on transactional databases with the time constraint to generate rules for environmental systems. First, they addressed quantitative association rule problem for a transaction database by applying K-means clustering and mapped them into Boolean association rules. The rule then discovered not only showed the occurrence of red tide and its relation with other attributes but also the time period of occurrence. However, this approach does not automatically extract interesting rules.

Temporal association rule mining discovers rule within a given timeframe only. However, we want to see temporal relationships where an occurrence of one unusual event is linked to other unusual events happening simultaneously or after a certain period of time, i.e. a delayed effect. For e.g. snowfall at night could cause busy traffic during next morning commute. This motivated us to look into related works on delayed correlation. Yamtani et. Al. (2014) used Delayed Correlation Analysis (DCA) to analyze the software evolution with the assumption that change in one variable during certain time period will affect other variables after some time delay. Also, the

correlation coefficient of 0.7 was used as the threshold to test for no correlation. Liang et al. (2015) used Generalized Cross Correlation (GCC) method on infrasound signal to estimate the time delay. In the paper, researchers have compared the performance of basic cross-correlation with GCC related methods Roth processor, SCOT, and PHAT and also with improved window function which performed better of all.

In our approach, we employ the concept of overlap and proximity to discover direct and time-delayed relation. We use anomalous clusters discovered in all domains to find these relations. If an anomalous cluster of one domain is overlapping with the other, then we identify direct relations for them. Otherwise, we check for proximity between anomalous cluster sets and identify the relation between those domains after shifting one domain by time-delay width δ .

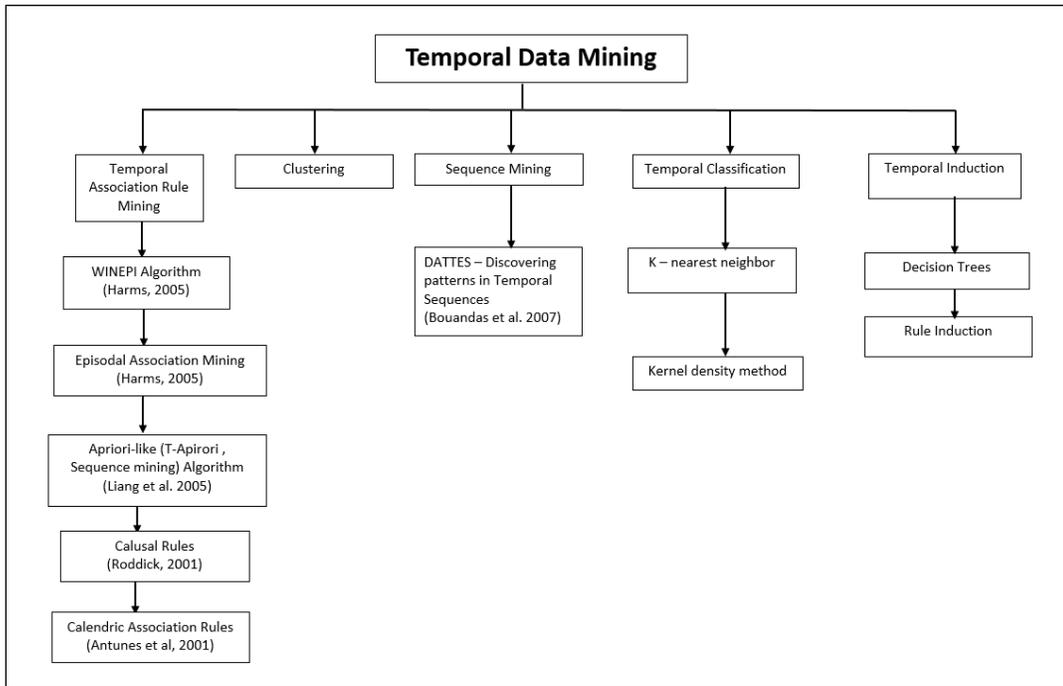


Figure 3: State of the art temporal mining techniques

Chapter 3: Methodology

In this chapter, we elaborate upon the methodology for the multi-domain temporal association. Our goal is to discover and quantify the temporal association between multiple distinct time-series domains using anomalous windows captured in each domain. To do so, we first discretize all domains into n number of bins using Equal Frequency method. Since data discretization segregates data into smaller sections, scan statistic can discover anomalies well because it can discover anomaly based on the normal/anomalous range for a smaller section rather than generalizing the range for the entire data. After binning, we discover anomalous windows in each bin for individual domains using scan statistic. For each bin, we check if a pair of anomalous time windows, with one anomalous window from each domain, has some overlaps or not. Overlap [Definition 1] indicates a direct linkage between two domains. For a set of domains with significant overlaps between anomalous windows pairs, we do further mining to get relations between them using the framework we proposed for temporal association rule mining. We also check for delayed correlation for sets of domains with less or no overlaps if their anomaly pairs are within the proximity [Definition 2]. For the set of anomalies within the proximity, further evaluations are done to see if there is a delayed correlation between those domains. Correlation indicates some link between these domains and to validate the link, we again perform association rule mining.

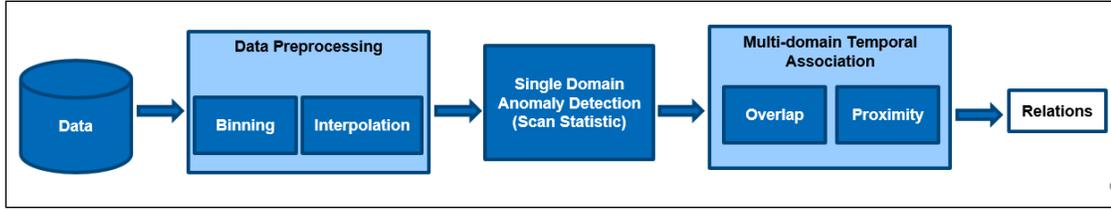


Figure 4: Multi-domain Temporal Association Process

Figure 4 above represents the multi-domain temporal association process. As shown in the picture, the first step is the data preprocessing in which we use binning and interpolation. The second step in anomaly detection for which we use scan statistic. The final step is the temporal association in which we employ the concept of overlap and proximity and discover relations between multiple distinct domains. Our methodology comprises of four major steps: 1. Discovery of anomalies, 2. Association of overlapping anomalies from various domains, 3. Checking for time-delayed relations, and 4. Quantifying the accuracy of multi-domain anomalies and relations.

S	<i>Spatial region</i>
T	<i>Total width of a time series data</i>
D	<i>Temporal domain</i>
A	<i>Anomalous window</i>
A_{D1}	<i>Anomalous window in domain D1</i>
A_{D1}^i	<i>i^{th} anomalous window in domain D1</i>
δ	<i>Time-delay coefficient</i>
$R_{D1 \Rightarrow D2}$	<i>Relation between domain D1 and D2</i>
P	<i>Proximity</i>
O	<i>Overlap</i>

tr_i	I^{th} transaction
t_{DI}^i	I^{th} time event in domain DI

Table 2: Summary of Terminologies used

3.1 Data Preprocessing:

Data transformation is paramount for any data mining process and it is not just hard but almost impossible to find a clean data ready for use. Data is transformed in order to gain computational efficiency. The process may include reducing the number of variables by removing variables with low information, removing noise and outliers, taking care of missing values, aggregation, discretization and so on. Data transformation is particularly of high importance in the time series data while dealing with multiple domains because the domains that are being used may be in different time granularity. Data transformations such as aggregation and discretization can be useful for such cases.

To facilitate our discovery in the data with missing values, we also evaluated state of the art algorithms for interpolation for multivariate temporal data.

3.1.1 Temporal Interpolation

One of the most common problems data scientists often have to deal with is data with missing values. Due to various reasons, such as instrument failure or lack of accessibility while recording values, data with missing values gets generated. Many data mining tools do not work well with such data. They either won't accept such data for processing or would produce a poor result after processing. So, we need to clean such data. To clean the data one may just delete the row with missing values or use

imputed values. But both options have some associated cost. Deleting data means losing data and losing information along with it. And imputing data always arises a question of if the imputed value is close enough to the original value. So, in this thesis, we did some research on interpolating time series data to find the best possible method to interpolate missing values.

Interpolation is a common data transformation technique used during the data preprocessing phase of data mining. It is basically used for the missing values problem or different time resolution problem in data. For a single variable problem, there are many straightforward and efficient interpolation techniques, such as linear interpolation or spline interpolation. However, these techniques don't perform very well as they aren't robust enough to capture the complexity of a multivariate missing data problem. We analyzed a few very advanced interpolation techniques that are not only robust enough to capture the complexity of a multivariate temporal data but also can capture the degree of interrelatedness between those variables.

Multivariate Imputation by Chained Equation (MICE) is an interpolation technique used for complex multivariate data, i.e. dataset having missing values in more than one attribute. The model for imputation used in this technique employs a conditional density based variable by variable iteration for each variable with missing values. We compared three variations of MICE techniques, and they are predictive mean matching (PMM), Norm, and Mean.

MICE(PMM): A semi-parametric method that can preserve non-linear relations. PMM can preserve non-linear relations in spite of the structural part of the imputation model being wrong. It is a very fast method and very efficient for a variable with many categories because PMM can impute variables as numeric variables. PMM only imputes observed values preserving the original categories of the (Buuren et al., 2011).

MICE(Norm): A simple and efficient method that uses the linear interpolation technique. For the model with residuals near to normal, it is fast and performs really efficiently.

MICE(Mean): Simply generates the mean as an imputed value and is a bad strategy (Buuren et al., 2011).

EOF: Another approach suggested by Zavala-Hidalgo et al. (2003) that uses a complex empirical orthogonal function methodology to capture moving patterns in the data. To apply this method, we first perform a Hilbert transform in the time domain.

During our study, we found that MICE Norm performed slightly better than MICE PMM, MICE Mean and EOF performed poorly. For our study, we used Premature Death and Population variable from MATCH data, and since there is a gradual change in population and death, and MICE Norm uses Linear Interpolation, MICE Norm performed better.

3.2 Discovery of anomalies

Discovering anomalies is the first step of our approach. Temporal data constitutes a series of events recorded at certain time intervals. Our goal in this step is to capture points or subsequences of events that are not normal with respect to the others. We

believe these unusual series of events often contain interesting knowledge. Hence, we try and capture these anomalous windows from all domains being analyzed and mine the knowledge extracted from them for further analysis. The process of anomaly detection is represented in Figure 5. Let's take a hypothetical example from the transportation application domain where the number of road accidents at a specific location is being recorded for each day, and 2-4 accidents per day is the normal and expected range. Time periods with extremely high or low number of accidents indicate unusual phenomena happening and this is where the knowledge lies. Once all such time periods are identified, we then find time periods representing unusual activity in other domains such as inclement weather conditions or, poor light conditions which are more likely to impact the number of accidents. These time periods of unusual activity are captured using the anomaly detection technique.

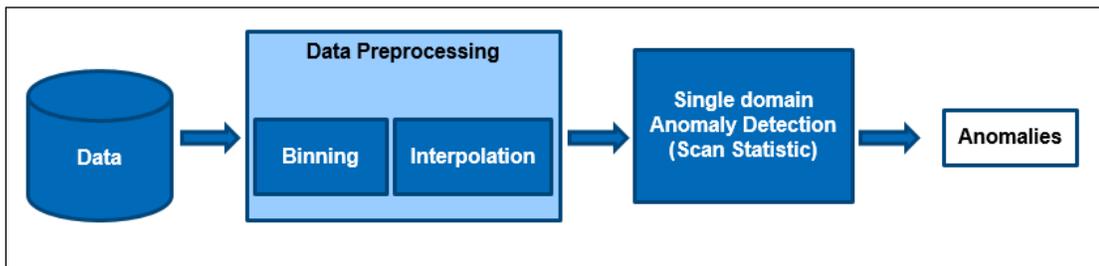


Figure 5: Anomaly detection process

Two data mining approaches that can be used for anomaly detection are, supervised learning methods where normal patterns are known beforehand and unsupervised learning method where normal patterns are derived based on the characteristics of the sample (Golmohammadi, 2015). As time series of complex domains do not follow a linear stochastic process or any regular pattern, distinguishing normal data points or

sequences from anomalies is particularly difficult (Golmohammadi, 2015). Our approach follows the unsupervised learning method using Scan Statistic, which uses local statistics to quantify the discovered window of a set of unusual points (Janeja, 2013). Scan statistic can be used to test pure randomness or presence of any cluster in a one-dimensional point process (Kulldorff, 1997). Kulldorff, 1997 developed the scan statistic to analyze geographical anomalous clusters (Kulldorff, 2001). Scan statistic is a group-based anomaly detection technique that discovers clusters of anomalies unlike point-based anomaly detection technique like FFT. It is an advanced statistical technique that can also be used for spatial, spatial-temporal or purely temporal data analysis. In our experiment, we used the purely temporal technique as we are analyzing time series data for a particular location. The first thing we do before using Scan Statistic was, discretizing the data into n number of bins. Data discretization improves the performance as it groups data together. We used the Equal Frequency binning technique, which is a very efficient yet simple and a straightforward technique. Then for each bin, we use Scan Statistic for anomaly detection.

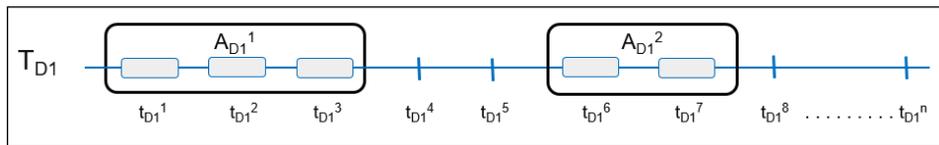


Figure 6: Anomalies in time series domain D1

Here, in Figure 6 we can see that for a first time-series domain $T_{D1} = \{t_{D1}^1, t_{D1}^2, \dots, t_{D1}^n\}$, where D1 represents the first time-series domain and t_{D1}^n is a time event recorded at time n, a set anomalous windows is represented as $A_{D1} = \{A_{D1}^1, A_{D1}^2,$

.... A_{DI}^i }, where $A_{DI}^i = \{t_{DI}^{n-p}, t_{DI}^{n-p+1}, \dots, t_{DI}^{n-q}\}$ is i^{th} anomalous window of the first domain and A_{DI}^i contains a subsequence of time events between t_{DI}^1 and t_{DI}^n .

3.3 Association of overlapped anomalous windows

Once all the anomalous windows for every distinct domain are discovered, the next step is to discover and quantify relations between these domains using the anomalous windows. The technique we follow for discovering the relation is a novel approach which is specific to temporal data. In our approach, for a given fixed spatial region S , we take a set of anomalous windows from all distinct time-series domains and use association rule mining to discover the relation between these domains. Before applying the algorithm, we first check for the number of overlaps between anomalous windows. If more than 50% of anomalous windows pairs are overlapped, then we move forward with our algorithm. Otherwise, we check for delayed correlation for those pairs.

DEFINITION 1: *[Overlap] Let t_x and t_y be time windows from domain x and y respectively. For time windows $t_x = \{t_x^1, \dots, t_x^n\}$ and $t_y = \{t_y^1, \dots, t_y^m\}$ overlap O_{xy}^i between t_x and t_y exists if both time windows have at least one identical time event i.e. $t_x^n = t_y^m$.*

Overlaps between anomalous time windows from two distinct domains mean some unusual activities happening in those domains during the same time period as shown in Figure 7. We assume that overlap indicates the direct relation between these distinct domains. However, overlaps can also occur due to a coincidence. So, to avoid such overlaps by chance we set a threshold for the number of identical time events in an

anomalous time windows pair and the number of bins with overlapping anomalous time windows. For a pair of anomalous time windows in a bin with anomalous time window from each domain, if more than 50% of total time events in each anomalous time windows are identical then they are said to have an overlap. If more than 50% of total number bins have anomalous time windows pairs with overlaps, then a set of domains are said to have significant overlaps. We only focus on sets of domains with significant overlaps and using association rule mining we discover relations between them and quantify these relations using support, confidence and lift. Otherwise, we do further analysis for time delayed relation.

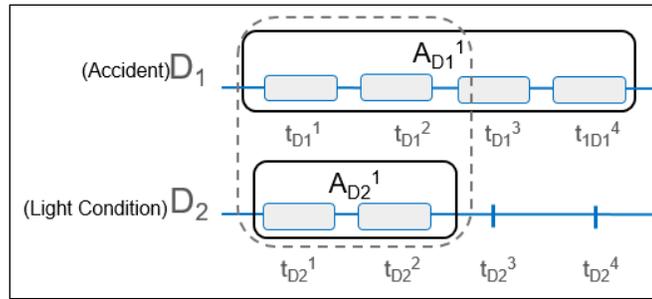


Figure 7: Anomalous time window overlap

DEFINITION 2: [Proximity] For n number of bins, let us take a pair of anomalous time windows with t_x and t_y , where t_x and t_y are anomalous time windows in the n th bin from domain x and y respectively. Let d_{xy} be the distance between t_x and t_y . Proximity P_{xy} is defined as the threshold used to determine the nearness between two time windows, t_x and t_y . It is calculated as $P_{xy} = T/(n*2)$, where T is the total number of time events in either domain and $T = T_x = T_y$ and n is the number of bins. Time window t_y is said to be in proximity with respect to t_x if $P_{xy} > d_{xy}$.

Time windows within proximity are considered to be neighbors. If no overlaps or overlaps in less than half of anomalous windows pairs are found, then we check if those pairs are within proximity or not. Based on the existence of proximity, we check for the delayed relation for the set of domains.

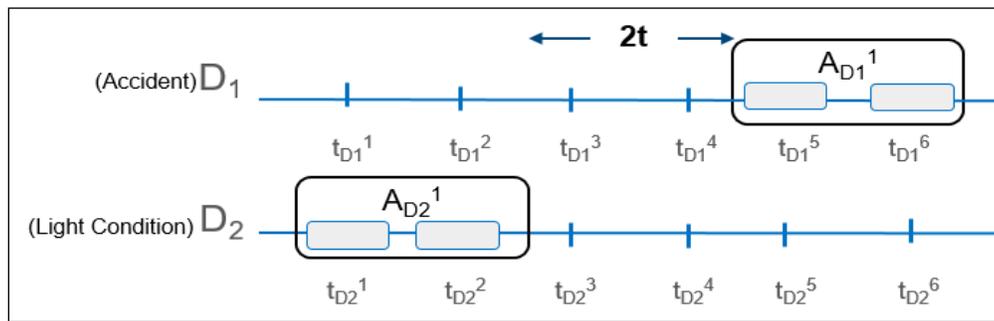


Figure 8: Proximity of $2t$

As we can see in Figure 8, anomalous windows A_{D1}^1 is said to be within proximity with respect to A_{D2}^1 if proximity, $P \geq 2t$.

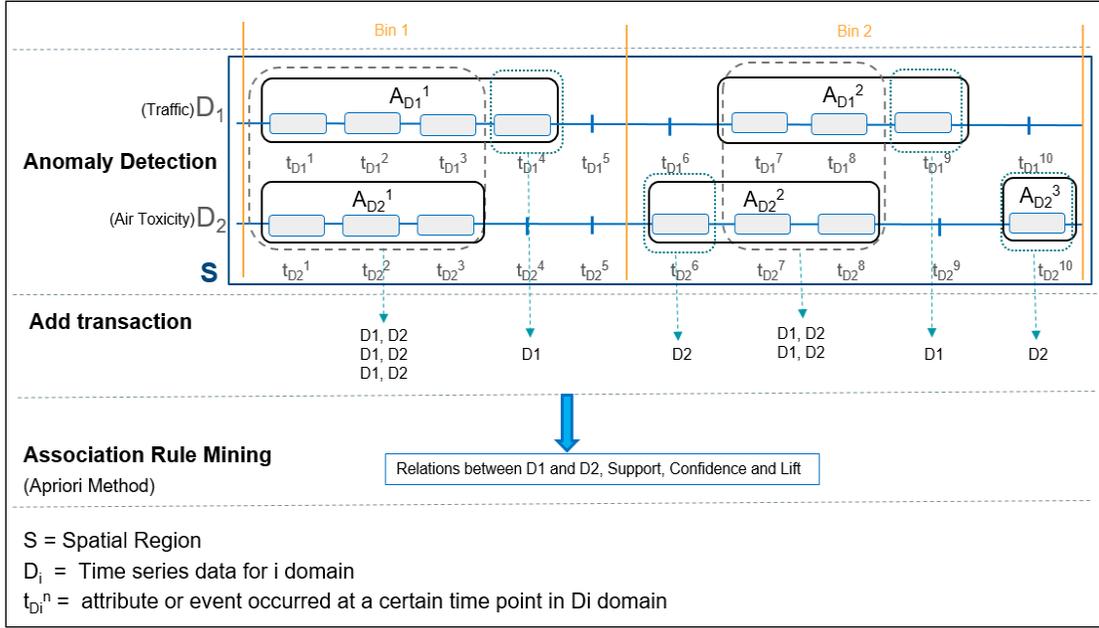


Figure 9: Multi-domain anomaly association framework

Figure 9 illustrates the technique of discovering multi-domain relation [Definition 3] where $A_{D_i}^x$ represents anomalous windows and dotted lines represent overlaps. For the time-series domain, traffic, we have anomalous windows, $A_{D_1}^1 = \{t_{D_1}^1, t_{D_1}^2, t_{D_1}^3, t_{D_1}^4\}$ and $A_{D_1}^2 = \{t_{D_1}^7, t_{D_1}^8, t_{D_1}^9\}$, where $t_{D_1}^n$ is an unusual time event recorded at time n . For another time series domain, air toxicity, we have anomalous windows, $A_{D_2}^1 = \{t_{D_2}^1, t_{D_2}^2, t_{D_2}^3\}$, $A_{D_2}^2 = \{t_{D_2}^6, t_{D_2}^7, t_{D_2}^8\}$ and $A_{D_2}^3 = \{t_{D_2}^{10}\}$. We can see that anomalous windows for these domains are overlapped at t^1, t^2, t^3, t^7 , and t^8 . Next, we generate a transaction where anomalous temporal events are treated as a transaction and domains with an anomaly in those temporal events are treated as items in a normal transaction as shown in Figure 10. Then we use the Apriori algorithm to compute the relation, support, confidence and lift.

t1	D1, D2
t2	D1, D2
t3	D1, D2
t4	D1
t5	
t6	D2
t7	D1, D2
t8	D1, D2
t9	D1
t10	D2

Figure 10: Sample transaction file used for association

DEFINITION 3: [Multi-domain Relation] For m number of domains $D1, D2, \dots, Dm$, let's take two domains $D1$ and $D2$. Each domain is discretized into n number of bins. A set of anomalous time windows in domain $D1$ and $D2$ is represented by A_{D1} and A_{D2} where $A_{D1} = \{A_{D1}^1, A_{D1}^2, \dots, A_{D1}^i\}$ and $A_{D2} = \{A_{D2}^1, A_{D2}^2, \dots, A_{D2}^j\}$. Each anomalous time windows consists of anomalous time events represented as t_{Dm} . For a set of anomalous time windows in n th bin S_{D1D2}^n , we check for overlap O_{D1D2}^n between anomalous time windows of $D1$ and $D2$. If overlap in more than 50% of bins is found, then we identify direct multi-domain relations $R_{D1 \Rightarrow D2}$ by using Apriori method. Otherwise, to identify delayed relations first we check if there is proximity P_{D1D2}^n between anomalous time windows of domain $D1$ and $D2$ in each bin. If proximity in all set of anomalous windows found, then using cross-correlation, we check for correlation between anomalous time windows in each set. If significant correlation is found, then we identify the time lag with highest correlation and shift delayed domain with that time lag. Then, using Apriori method we identify relations $R_{D1 \Rightarrow D2}$ between $D1$ and $D2$.

Our technique of association is shown step by step in Algorithm 1. As an input for our algorithm, we use a list of temporal anomalous windows from each domain. The first step is to calculate overlaps in anomalous windows pairs, which is shown in line 1. As we are interested in direct and time-delayed relations, we compute the number of pairs of anomalous windows with overlaps, which indicates a direct relation. For the set of domains with at least 50% of overlapping pairs, we further continue the analysis with our algorithm. Otherwise, we check to see if those pairs have delayed correlation, and if they do then we use our association algorithm to get relations. In line 3, for each pair of anomalous windows, we check if there are time overlaps and increase the counter. In line 5, we check if the number of bins with overlapping anomalous window pairs is more than 50% or not. In line 6 to 11, we add a transaction for every overlapping and non-overlapping anomalous time events. Finally, in line 13, we use the association rule mining and calculate confidence, support, and lift. As an additional measure of validation, in line 14 we also calculate the correlation for the anomalous time window pairs.

Algorithm 1: Multi-domain Temporal Anomaly Association Algorithm

Input: Set of anomalous windows $A_{D1} = \{A_{D1}^1, A_{D1}^2, \dots, A_{D1}^n\}$, $A_{D2} = \{A_{D2}^1, A_{D2}^2, \dots, A_{D2}^n\}$ for two domains

Output: relations $R_{D1 \Rightarrow D2}$ between domain D_1 and D_2 , support S , confidence C and lift L

Pseudo code:

```

1: for each pair of anomalous windows in a bin,  $A_{D1}^n$  and  $A_{D2}^n$ , where  $A_{D1}^n = \{t_i, t_{i+1}, \dots, t_j\}$  and  $A_{D2}^n = \{t_k, t_{k+1}, \dots, t_l\}$ 
2:   Check for overlaps of time events in the anomalous windows pair:
   if  $t_i = t_k$ , overlap found
3:     Counter++
4:   end if
5:   if Counter > 50 % of anomaly pairs
6:     for each time event in overlapped time subsequence
7:       Add a transaction,  $tr_i = D_1, D_2$ 
8:     end for
9:     for each non-overlapped time events in domain p and q
10:      Add a transaction,  $tr_i = D_1$  for domain 1,  $tr_i = D_2$  for domain 2
11:    end for
12:  end for
13:  Apply Apriori on transaction  $tr_i$ 
14:  Calculate correlation.
15: end if

```

3.4 Time-delayed relation

As we already discussed above, unusual activity in a domain can be an impact of some unusual activity happening in another domain. However, impacts of one domain to the other aren't always simultaneous. There can be some associated time-delayed impacts. For example, heavy snow will significantly impact the traffic for the following few days, the oil spill in an ocean can impact aquatic lives as well as nearby wild lives after a certain amount of time, and so on. For domains mentioned in the previous examples, there might not be an overlap of anomalous time windows regardless of significant

impacts of one domain on the other. We will lose the knowledge of association in such a scenario if a time delay factor is not considered and evaluated. So, in this section, for each bin, we check for delayed correlation between anomalous time windows if they are within the proximity. If a correlation is found, then we follow the same steps as Algorithm 1 for discovering relations. Figure 11 illustrates the process of discovering relations for time-delayed impacts. We can see that no anomaly in domain D1 has overlapping time events with anomalies in domain D2 which implies that there is no direct relation between these two domains. However, there still can be a time-delayed relation. So, in the second step of Figure 11, we check for proximity. We can see that anomalous windows pairs, (A_{D1}^1, A_{D2}^1) and (A_{D1}^2, A_{D2}^2) are within proximity of $P = 2t$. Next, we check if anomalous windows in a bin have some correlation by using cross-correlation with the lag of δ , then we identify the time lag with maximum correlation δ_{max} and shift a domain with the δ_{max} value. We then create transaction and use Apriori method on that transaction like we did in Algorithm 1.

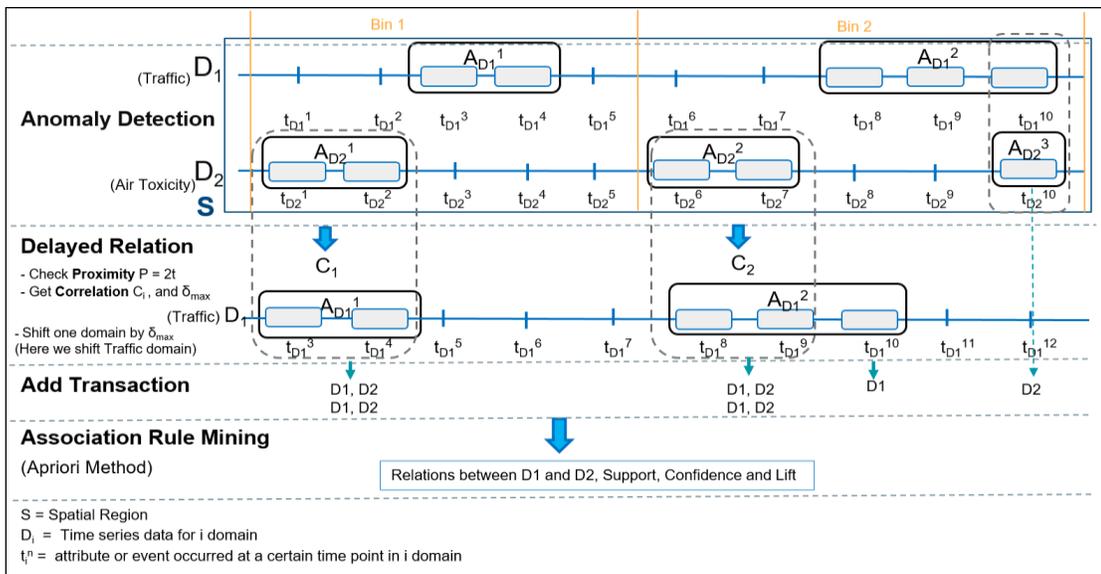


Figure 11: Time-delayed multi-domain anomaly association

In line 1, first, we compute the proximity and check if time windows in anomalous window pairs are within the defined proximity. In line 2, for each pair, we compute the average time difference between anomalous time windows in a pair. If the average time difference is lower than proximity, then we move forward to check for correlation. In line 6, we compute the cross-correlation between domains. If a correlation is found, then that indicates some delayed relation between these domains. So, we identify the time lag (δ) with highest correlation δ_{\max} then in line 7, we shift one domain by δ_{\max} , then for each pair of anomalous windows, we compute correlation. And to discover and quantify the relations, in line 18, we apply association rule mining.

Algorithm 2: Multi-domain Temporal Time-delayed Anomaly Association

Input: set of anomalous windows $A_{D1} = \{A_{D1}^1, A_{D1}^2, \dots, A_{D1}^n\}$, $A_{D2} = \{A_{D2}^1, A_{D2}^2, \dots, A_{D2}^n\}$, time delay coefficient δ

Output: relation $R_{D1 \Rightarrow D2}$ between domains D_1 and D_2 , correlation, support S , confidence C , lift L

Pseudo code:

- 1: Compute proximity threshold: Proximity $P = T / (x * 2)$, $x = \text{no. of bins}$
- 2: **foreach** set of anomalous windows pair
- 3: Calculate the average time gap t_{avg} , between those anomaly pairs
- 4: **end for**
- 5: **if** (time gap $t_{\text{avg}} < \text{proximity}$)
- 6: Use cross-correlation to find time lag (δ) with highest correlation δ_{\max}
- 7: Shift one domain by width of δ
- 8: **foreach** pair of anomalous windows
- 9: Compute the correlation between the pair
- 10: **foreach** overlapped time events

```

11:             Add a transaction,  $tr_i = D_1, D_2$ 
12:         end for
13:     foreach non-overlapped time events of domain p and q
14:         Add a transaction,  $tr_i = D_1$  for domain 1,  $tr_i = D_2$ 
           for domain 2
15:     end for
16: end for
17: end if
18: Apply Apriori method on  $tr_i$ 

```

3.5 Accuracy Measures

We quantify the results and measure the performance of our algorithms. To quantify how well our algorithm did, we calculate support, confidence and lift using the Apriori algorithm. For overlapped pairs of windows, we also calculate correlation to see if the pairs are actually related. We also measure the accuracy of the anomaly detection technique, to see how accurately all the anomalies are being identified by scan statistic. To measure the accuracy, we use a contingency table as shown in Figure 12. We count all the True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) anomalous time events picked up by scan statistic and then we calculate precision, recall, and accuracy. Formulas for calculating precision, recall, and accuracy are listed below in Table 3.

Measures	Formula
Precision	$TP / (TP + FP)$
Recall	$TP / (TP + FN)$

Accuracy	$(TP + TN) / (TP + TN + FP + FN)$
----------	-----------------------------------

Table 3: Precision, Recall and Accuracy formula

	Predicted Condition		
True Condition		True	False
	True	TP	FN
	False	FP	TN

Figure 12: Contingency Table

We also use Piecewise Aggregate Approximation (PAA) for performance evaluation. It is a dimensionality reduction method widely used for time series data. Here, we use this technique as a data visualization to compare the values generated by PAA with anomalies discovered by scan statistic.

Chapter 4: Experiment and result

In this section, we discuss the tools, libraries, and datasets that were used for the experiment. We present the result and accuracy we got after the discovery of anomalies, association, and correlation for each dataset. We also present the performance evaluation of interpolation methods and identify the best method.

4.1 Software and packages specification

For our experiment, we built the web-like application using R language which comes with wide variety of packages that can be installed and used in the form of a library. R is basically a statistical analysis software but has a wide range of mathematical, visualization, data analysis, and many more packages. Shiny is a very interesting package in R which allows us to build a graphical web-like application, where we can automate many steps. R also has a package called ‘rsatscan’, which interacts with the SatScan software and makes it possible to do everything from R. ‘Shiny’ and ‘rsatscan’ are key packages that have a very prominent role in our experiment. We also used ‘mice’ and ‘sinkr’ [22] packages for interpolation methods and ‘jmotif’ package for PAA. And finally, for finding associations using Apriori method, we used ‘arules’ package.

4.2 Datasets

We used two multi-domain real-world datasets MATCH (Mobilizing Action Toward Community Health), NJDOT (New Jersey Department of Transportation) and weather

data to experiment and validate our approach. We also used a synthetic data to allow us to measure the performance of our approach.

Here, we discuss each dataset that was used and we also present the result found from the anomalous windows associations for each of them. Also, we present the accuracy and performance analysis outputs and for real world data.

4.2.1 MATCH Health Ranking data

MATCH is one of the real-world datasets we used for our temporal anomaly association approach. We analyzed the premature death and excessive drinking to find the relation between these two domains. MATCH dataset is provided by County Health Ranking and Roadmaps program, which is a collaboration between the Robert Wood Johnson Foundation and the University of Wisconsin Population Health Institute. This program has been ranking every county in the US every year by measuring various factors that have an impact on a county's health. They measure how healthy a county is based on the mortality (length of life) and morbidity (quality of life). In general, the health of a county is influenced by factors such as health behaviors, clinical care, social and economic factors, and physical environment. Based on all these factors, counties are ranked every year and the approach is shown in Figure 13.

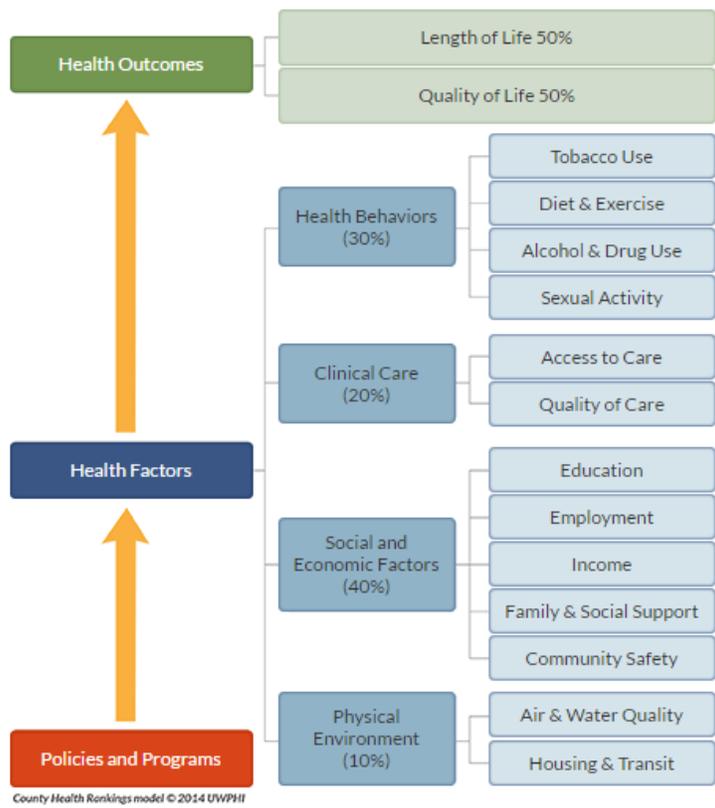


Figure 13: Community health ranking approach

This dataset consists of various factors listed in Table 4 below. and data from the year 2010 to 2016 is publicly available. This ranking is done every year; hence, time resolution is a year. For our experiment, we used the premature health and excessive drinking domain data for Bergen county from the year 2010 to 2016 to find the relation between these two domains. Since we are concerned with temporal data, we only had 7 rows of yearly data.

Health Behaviors	adult smoking, adult obesity, food environment index, physical inactivity, access to exercise opportunities, excessive drinking, alcohol-impaired driving deaths, teen births, sexually transmitted infections
-------------------------	--

Clinical Care	uninsured, primary care physicians, dentists, mental health providers
Social and Economic Factors	high school graduation, some college, unemployment, children in poverty, income inequality, social associations, children in single-parent households, violent crime, injury deaths
Physical Environment	air pollution-particulate matter, drinking water violations, severe housing problems, driving alone, long commute—driving alone

Table 4: Measures used to rank county’s health condition

To discover the anomaly, scan statistic requires data to be in a certain format. It needs a population file which consists of the total population for the domain being use and the case file, which has the number of impacted cases to be analyzed. The name and description of the files we used for our experiment for MATCH data are listed below in Table 5.

Domain Dataset	Description
Premature_death_Bergen.csv	Number of premature deaths in Bergen county
Excessive_drinking_Bergen.csv	Number of excessive drinking in Bergen county
Bergen_pop.csv	Total population in Bergen county

Table 5: List of files used for MATCH data

4.2.2 NJDOT data

New Jersey Department of Transportation maintains the data of crash, driver, vehicle, occupant, and pedestrian from the year of 2001 to current date. Road accidents can be caused for a various reason, such as weather condition, a section of the road, a condition of the driver or the vehicle hence NJDOT records accidents data along with casualty,

road conditions, light condition and so on every day for all counties of New Jersey state. Variables recorded by NJDOT for a crash are listed in Table 6 below.

Location	County Code, Municipality Code, County Name, Municipality Name, Crash Location, Location Direction, Cross Street Name, Latitude, Longitude
Time	Year, Crash Date, Crash Day of Week, Crash Time
Casualty	Total Killed, Total Injured, Pedestrians Killed, Pedestrians Injured, Severity
Driver	Cell Phone in Use Flag, Alcohol Involved
Road	Intersection, Route, Route Suffix, SRI (Standard Route Identifier), Mile Post, Road System, Road Character, Road Surface Type, Road Divided By, Distance to Cross Street, Direction from Cross Street, Is Ramp, Ramp To/From Route Name, Ramp To/From Route Direction, Posted Speed, Posted Speed Cross Street
Environmental condition	Surface Condition, Light Condition, Environmental Condition
Others	Department Case Number, Police Department Code, Police Department, Police Station, HazMat Involved, Crash Type Code, Total Vehicles Involved, Temporary Traffic Control Zone, Other Property Damage, Reporting Badge No

Table 6: Categorized variables in NJDOT data

For our multi-domain anomaly association experiment, we used the Total Injured, Light Condition and Surface Condition domain of Bergen county of June 2014. The number of rows of the subset of data we used is 2502. Domain Total Injured has numeric values,

but Light Condition and Surface Condition has symbolic values which are listed in Table 7 and Table 8. The name and description of the files we used for our experiment for NJDOT data are listed below in Table 9.

Light condition	Light condition code
Day	'01'
Dusk	'03'
Night	'04','05','06' or '07'
Dawn	'02'
Unknown	'00'
Other	'99'

Table 7: Light condition codes

Surface condition	Surface condition code
Dry Surface	'01'
Wet Surface	'02' or '06'
Snow	'03'
Ice	'04'
Unknown	'00'
Other	'05', '07', '08' or '99'

Table 8: Surface Condition Code

Domain Dataset	Description
Total_injured_Bergen.csv	Number of people injured in an accident in Bergen county

Light_condition_Bergen.csv	Light condition during an accident in Bergen county
Surface_condition_Bergen.csv	Surface condition during an accident in Bergen county
Bergen_pop_.csv	Total population in Bergen county

Table 9: List of files used for NJDOT data

4.2.3 Synthetic data

In addition to the above two real-world datasets, we also used a synthetic dataset where we inserted few anomalies. Synthetic data is used as another measure of accuracy to see how well the known anomalous windows are being picked by the anomaly detection we are using and also to test if known relation between domains are being detected by the algorithm we proposed. The name and description of the files we used for our experiment for synthetic data are listed below in Table 10.

Domain Dataset	Description
Total_injured_Bergen_1.csv	Number of people injured in an accident in Bergen county
Light_condition_Bergen_1.csv	Light condition during the accident in Bergen county
Surface_condition_Bergen_1.csv	Surface condition during an accident in Bergen county
Bergen_pop.csv	Total population in Bergen county

Table 10: List of files used for synthetic data

4.2.4 Weather data:

This weather data is recorded hourly by the weather sensors at beaches along Chicago's Lake Michigan lakefront. The data is maintained and has been made publicly available by Chicago Park District. The data comprises of variables as timestamp, air temperature, humidity, wind direction, solar radiation and so on. The name and description of the files we used for our experiment for this weather data are listed below in Table 11.

Domain Dataset	Description
SolarRadiation.csv	Solar radiation recorded for Oak Street Station
Temperature.csv	Temperature recorded for Oak Street Station
Humidity.csv	Humidity recorded for Oak Street Station
POP.csv	Total population

Table 11: List of files used for weather data

4.3 Experiment and results

In this section, we present the results we got after applying our approach to the datasets mentioned above. Time resolution, time period, the number of data points and domains of each dataset is presented in Table 12. After binning a dataset, we get a set of buckets of datasets on which we applied our novel algorithm and discovered relations between multiple distinct domains in the dataset. As a final step, we quantify the relations found using confidence, support, lift, and correlation. We also used cross-correlation and PAA as external validation techniques to validate discovered relations.

Dataset	Domains	Number of rows	Time resolution	Time period
MATCH	premature death, excessive drinking	7	yearly	7 year
NJDOT	total injured, light condition, surface condition	2502	daily	1 month
Synthetic	total injured, light condition, surface condition	2502	daily	1 month
Weather	solar radiation, temperature, humidity	14690	hourly	698 days

Table 12: Details of Datasets Used

4.3.1 Data Preprocessing

In this section, we present the result we got for the comparative study of state of the art multivariate interpolation techniques for temporal interpolation.

4.3.1.1 Temporal Interpolation

Experiment for interpolation was also conducted using the R application with four Interpolation methods: MICE PMM, MICE Norm, MICE Mean, EOF. For the experiment, 5% (999 rows) of data from MATCH (Mobilizing Action Toward Community Health) dataset was used. Once the data is uploaded, application randomly removes few values of data to create a gappy data then the following interpolation methods were used to fill those missing values: MICE PMM, MICE Norm, MICE Mean, EOF. The bivariate approach was used to calculate the error for all interpolation methods.

Then accuracy was measured by subtracting interpolated values from original values for two variables (no. of death, total population). Then, values of both variables were normalized in the range of 0 to 1 and the average of those normalized variables was plotted using MS Excel. Standard Deviation (SD) of the average was taken and line of

2*SD and 3*SD was added to the plot. Data points above standard deviation line were considered to be False Positive. Then, using Predictive Positive Values (PPV) method, accuracy was calculated and accuracy of all methods was compared.

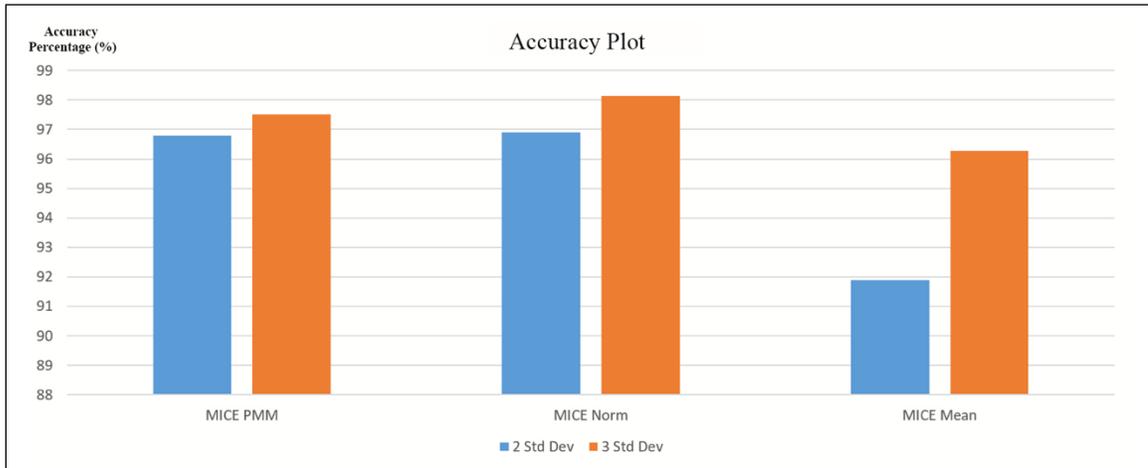


Figure 14: Interpolation techniques accuracy comparison plot

After the above experimentation, it was found that EOF method was not able to interpolate all the missing values. And MICE Norm method performed best with an accuracy of 96.89% for 2*SD and 98.13% for 3*SD which is shown in Figure 14.

4.3.2 Single domain anomalies

We discovered sets of temporal anomalous windows using the Purely Temporal analysis method and Discrete Poisson model in scan statistic.

4.3.2.1 Discovered anomalies in MATCH data

For MATCH data, we used the Premature Death and Excessive Drinking domain for our multi-domain anomalies association approach. We discretized data in both domains using Equal Frequency method into two bins. Then using the purely temporal technique

in scan statistic, we discovered anomalies in each bin which are listed below in Table 13 and shown in Figure 15.

Bin No.	Anomalies from Premature Death		Anomalies from Excessive Drinking	
	Anomalies	P-value	Anomalies	P-value
1	2011/1/1 – 2011/1/3	0.001	2010/1/1 – 2010/1/5	0.001
2	2015/12/30 - 2016/1/1	0.001	2012/12/30 – 2013/1/1	0.001

Table 13: Anomalies in domains of MATCH data

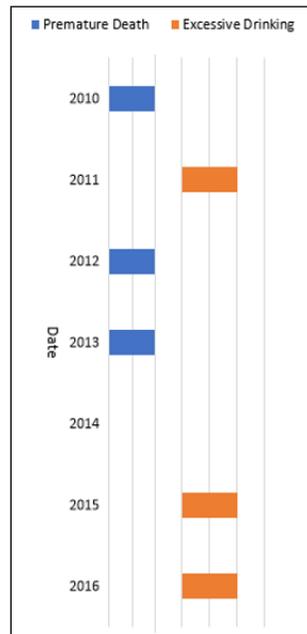


Figure 15: Anomalies in MATCH data

Since we only had 7 rows of data for MATCH, we chose to have two bins.

4.3.2.2 Discovered anomalies in NJDOT data

We used the same technique that we used for MATCH data, except we discretized our data into three bins using the same discretization technique, Equal Frequency. For NJDOT data, we used Total Injured, Light Condition and Surface Condition domains.

Temporal anomalies discovered for all domains are listed in Table 14 and presented in Figure 16.

Total injured		Light condition		Surface condition	
Anomalies	P-value	Anomalies	P-value	Anomalies	P-value
2014/6/1 – 2014/6/9	0.023	2014/6/4 – 2014/6/9	0.008	2014/6/10 – 2014/6/12	0.014
2014/6/13 – 2014/6/15	0.028	2014/6/13 – 2014/6/18	0.001	2014/6/13 – 2014/6/15	0.001
2014/6/19 – 2014/6/24	0.044	2014/6/19 – 2014/6/24	0.007	2014/6/19 – 2014/6/21	0.017

Table 14: Anomalies in domains of NJDOT data

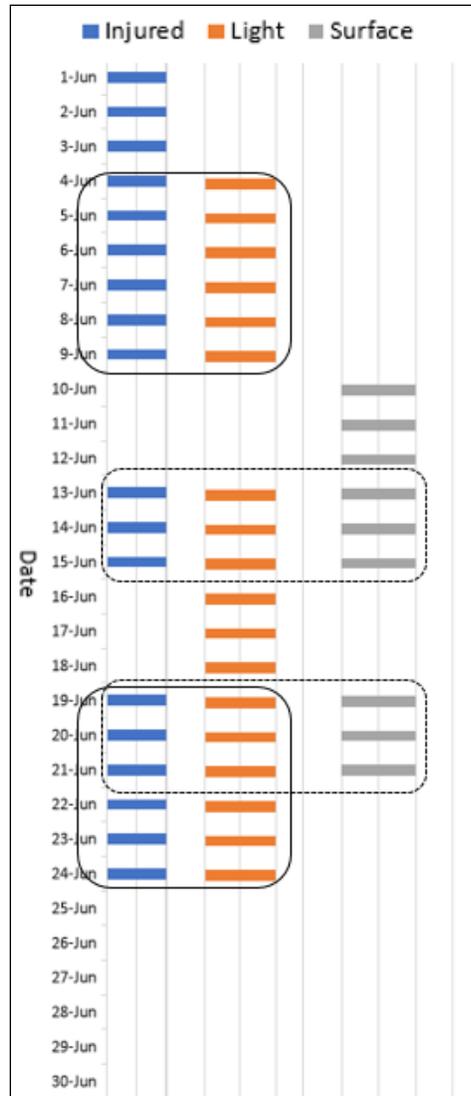


Figure 16: Anomalies in NJDOT data for all domains

4.3.2.3 Discovered anomalies in synthetic data

We also used a synthetic data as an additional set of data to experiment on and also as a means to test the performance of our temporal anomalies association approach. We inserted anomalous time windows by adding unusually high values in 5% of the NJDOT data to test if those anomalies get picked up by the scan statistic. We inserted three anomalous time windows in each domain. Inserted anomalous windows in every

domain are listed in Table 15 and anomalies discovered by scan statistic are listed in Table 16 and presented in Figure 17.

Total injured	Light condition	Surface condition
2014/6/1 – 2014/6/5	2014/6/1 – 2014/6/5	2014/6/1 – 2014/6/5
2014/6/13 – 2014/6/17	2014/6/13 – 2014/6/17	2014/6/13 – 2014/6/17
2014/6/22 – 2014/6/27	2014/6/22 – 2014/6/27	2014/6/22 – 2014/6/27

Table 15: Inserted anomalous windows in synthetic data

Total Injured		Light Condition		Surface Condition	
Anomalies	P-value	Anomalies	P-value	Anomalies	P-value
2014/6/1 – 2014/6/9	0.004	2014/6/4 – 2014/6/9	0.001	2014/6/1 – 2014/6/6	0.001
2014/6/13 – 2014/6/18	0.012	2014/6/13 – 2014/6/18	0.001	2014/6/13 – 2014/6/18	0.001
2014/6/19 – 2014/6/27	0.001	2014/6/22 – 2014/6/27	0.001	2014/6/22 – 2014/6/27	0.001

Table 16: Anomalies in synthetic data

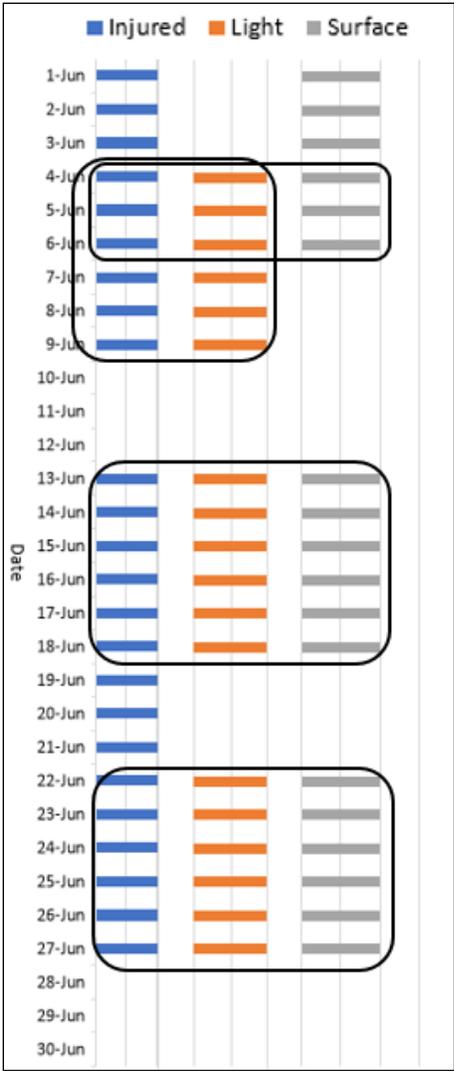


Figure 17: Anomalies in synthetic data for all domains

4.3.2.4 Discovered anomalies in weather data:

We used the same technique that we used for other data, we discretized our data into eight bins using the same discretization technique, Equal Frequency. For weather data, we used Solar Radiation, Temperature and Humidity domains. Temporal anomalies discovered for all domains are listed in Table 17 and presented in Figure 18.

Solar radiation		Temperature		Humidity	
Anomalies	P-value	Anomalies	P-value	Anomalies	P-value
2015/7/20 to 2015/8/6	0.001	2015/7/17 to 2015/8/3	0.001	2015/6/11 to 2015/6/22	0.001
2015/9/12 to 2015/9/26	0.001	2015/8/31 to 2015/9/17	0.001	2015/9/24 to 2015/10/6	0.001
2016/2/6 to 2016/2/20	0.001	2015/11/29 to 2015/12/16	0.001	2015/11/26 to 2015/12/13	0.001
2016/5/18 to 2016/5/29	0.001	2016/5/21 to 2016/6/1	0.001	2016/4/27 to 2016/5/2	0.001
2016/6/5 to 2016/6/19	0.001	2016/8/1 to 2016/8/18	0.001	2016/7/29 to 2016/7/31	0.001
2016/8/28 to 2016/9/14	0.001	2016/9/3 to 2016/9/20	0.001	2016/9/30 to 2016/10/2	0.001
2016/11/14 to 2016/11/22	0.001	2016/11/14 to 2016/12/1	0.001	2017/1/16 to 2017/1/27	0.001
2017/4/7 to 2017/4/21	0.001	2017/4/7 to 2017/4/21	0.001	2017/3/26 to 2017/3/31	0.001

Table 17: Anomalies in weather data

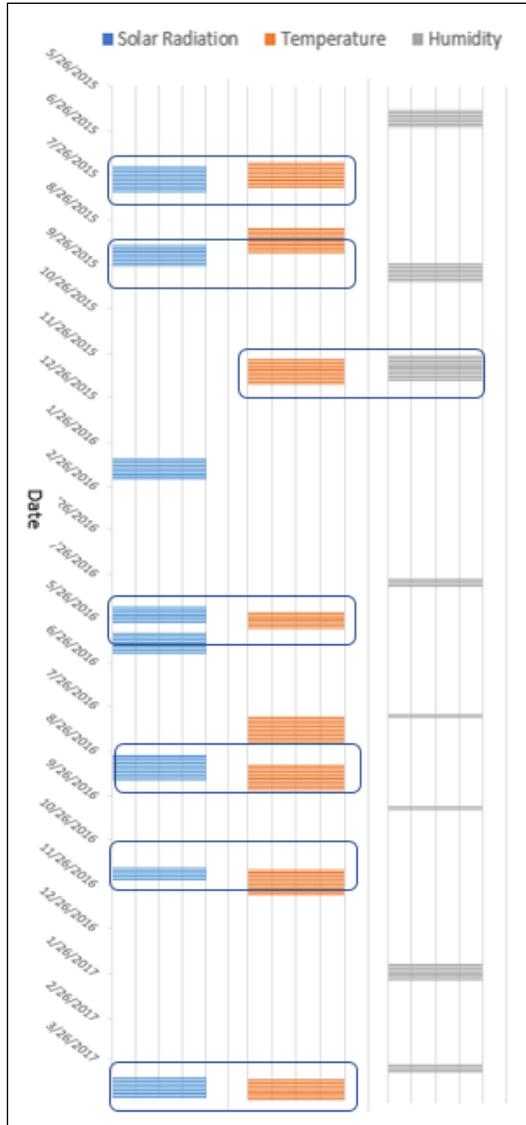


Figure 18: Anomalies plot for weather data

4.3.3 Temporal Anomalies Association

In this step, we present the results we got after applying our novel algorithm on the anomalies listed above to discover relations between domains in each dataset. We created a transaction for each set of domains and treated each domain in an anomalous time window as an item. Then using the available Apriori algorithm, we computed the

relation between those domains. We set the threshold for support = 0.5 and confidence = 0.75.

Since in the weather data we didn't have any overlaps between three domains, we analyzed weather data further for delayed relation in section 4.3.4.

4.3.3.1 Association in MATCH data

We used the Apriori method for anomalies found in Premature Death and Excessive Drinking domains and no relation was found between Premature Death and Excessive Drinking and it is an expected result as there are no overlaps between these two domains. We were constrained by the size of this data to approach the problem with other ways such as using a different subset of data or using a different number of bins.

4.3.3.2 Association in NJDOT data

Following the same technique as above and using Apriori method we discovered relations between Total Injured, Light Condition and Surface Condition domains which are listed in Table 18.

Relation Found	Support	Confidence	Lift
{Injured} => {Light}	0.60	0.8333	1.1574
{Light} => {Injured}	0.60	0.8333	1.1574
{Injured, Surface} => {Light}	0.24	1.0000	1.3888
{Light, Surface} => {Injured}	0.24	1.0000	1.3888
{Injured, Light} => {Surface}	0.24	0.40	1.11
{Surface} => {Injured}	0.24	0.67	0.926
{Injured} => {Surface}	0.24	0.33	0.926
{Surface} => {Light}	0.24	0.67	0.926
{Light} => {Surface}	0.24	0.33	0.926

Table 18: Support, Confidence and Lift in NJDOT data

In Table 18 above, in the first two relations we can see the mutually dependent relation between Light Condition and Total Injured with support = 0.6, confidence = 0.83 and lift = 1.157 which implies that Light Condition and Total Injured domain are strongly related. In the third relation, we can also see a relation between Total Injured, Surface Condition and Light Condition with support = 0.24, confidence = 1 and lift = 1.389, which implies that there is a relation between these domains because of lower support and higher confidence.

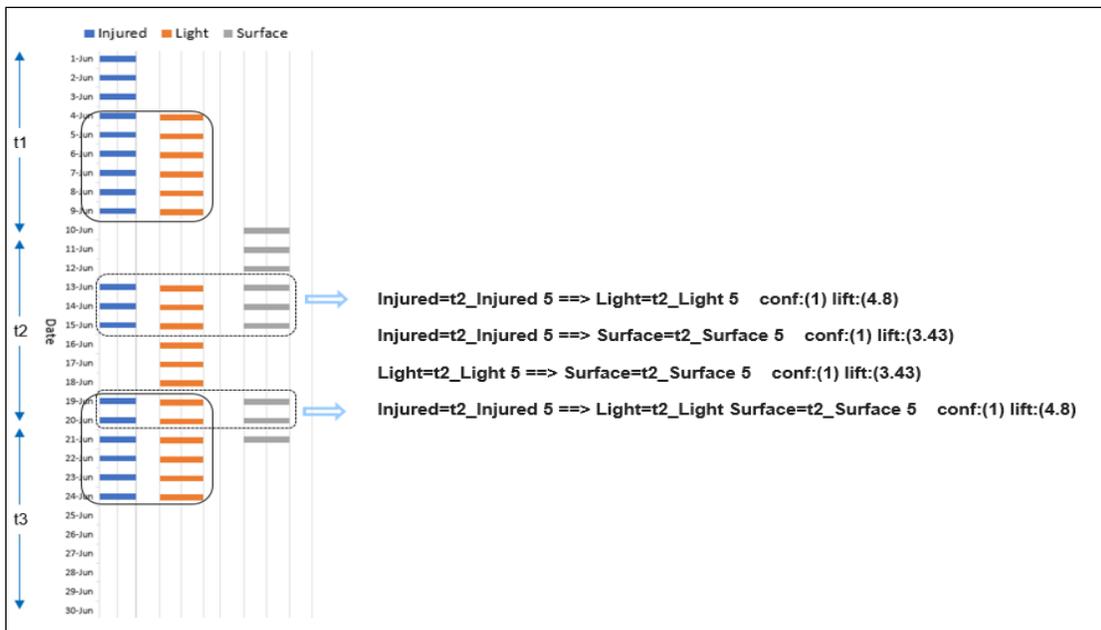


Figure 19: Relations with time information

We also added time information in the transaction to see which relations in which bin are being discovered. As we can see in Figure 19, we have marked the time events for bin 1 of a domain as t1_DomainName, i.e. for time events of the Injured domain in bin

1, we marked them as t1_Injured, in bin 2 as t2_Injured and so on. In Figure 18, we can see that we are getting strong relations for bin2. In the first relation, we have a relation between Total Injured and Light Condition domains with the confidence of 1 and lift of 4.8, and in the last relation, we have a relation between all three domains with the confidence of 1 and lift of 4.8.

4.3.3.3 Association in synthetic data

Relations discovered between Total injured, Light Condition and Surface Condition using Apriori method for synthetic data are listed in Table 19.

Relation Found	Support	Confidence	Lift
{Surface} => {Light}	0.625	0.833	1.111
{Light} => {Surface}	0.625	0.833	1.111
{Injured, Surface} => {Light}	0.625	0.833	1.111
{Injured, Light} => {Surface}	0.625	0.833	1.111
{Light, Surface} => {Injured}	0.625	1.000000	1.000000
{Surface} => {Injured}	0.750	1.000000	1.000000
{Injured} => {Surface}	0.750	0.750000	1.000000
{Light} => {Injured}	0.750	1.000000	1.000000
{Injured} => {Light}	0.750	0.750000	1.000000

Table 19: Support, Confidence and Lift in synthetic data

In Table 19 above, in the first two relations, we have the mutually dependent relation between Surface and Light Condition domain. In the third relation, we can see that there is a significant relation between Total Injured, Light Condition and Surface Condition with support = 0.625, confidence = 0.833 and lift = 1.111, implies that Total Injured, Light condition and Surface Condition domains are strongly related. This is an

expected result given that we inserted anomalies for same time periods in both of these domains.

The relations listed in Table 18 and Table 19 for NJDOT and synthetic data respectively shows that our novel algorithm was able to capture the relation between associated domains.

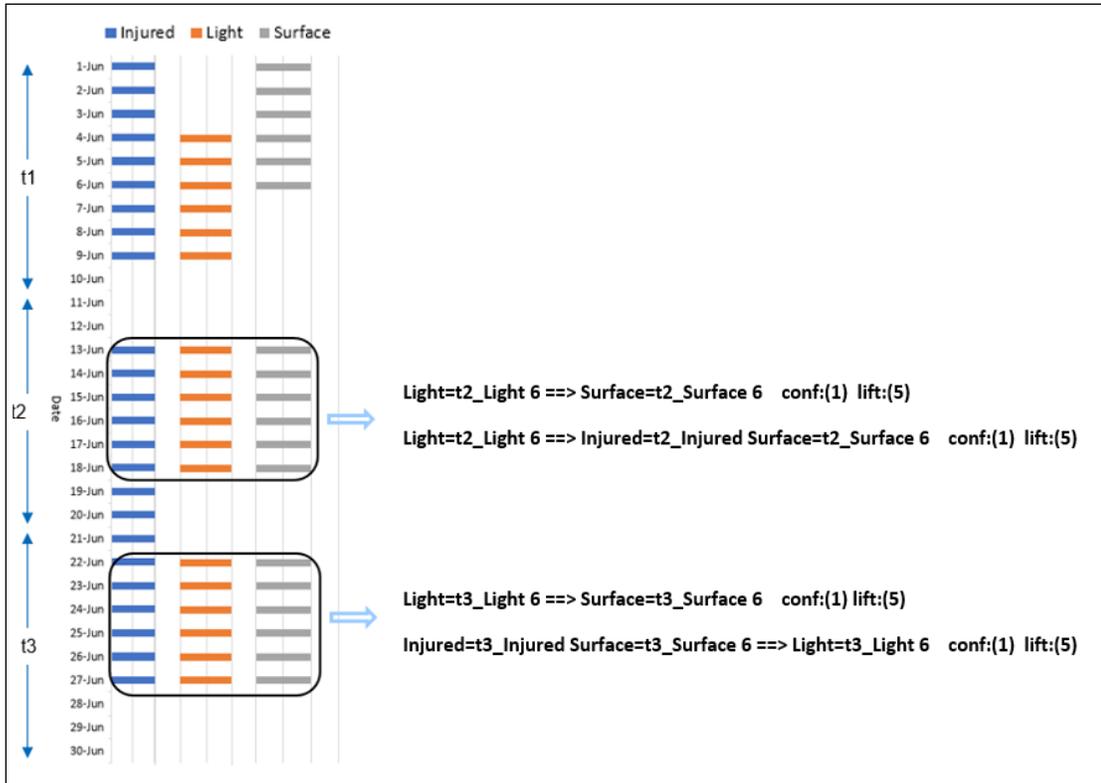


Figure 20: Relations with time information

In Figure 20 above, we can see the relations with time information. The first relation is between Light and Surface Condition domain in bin 2, and the relation is quite strong with the confidence of 1 and lift of 5. In the third relation, we have a relation from bin 3 for Light and Surface condition domain. Similarly, in the second and last relation, we can see the relation between all three domains for bin 2 and bin 3 respectively. If we

look at the figure on left side, we can see that there are significant overlaps between all three domains.

4.3.4 Time-delayed Association

To find out if the set of domains have a time-delayed relation, we check if anomalous window pairs of those domains are within proximity or not. If the set of anomaly pairs are within proximity, then we further analyze them to check for the delayed relation.

For MATCH data, we didn't find any overlap and for NJDOT and synthetic data, there are significant overlaps of anomaly pairs. For MATCH data, we were not able to proceed further with this step because we only had 7 records of data to work with, which is relatively very less for any data mining analysis. In NJDOT and synthetic data, significant overlaps and relation between all domains were found which indicates they have a stronger direct relation. So, we didn't analyze these data time-delayed relation.

However, in weather data, there was no overlap between three domains so we further analyzed this data for time-delayed relation. First, we computed the cross-correlation between each domain using the lag of δ . We used $\delta = 43$ because we are using one data with 698 days and binning it into eight bins, which makes the size of time events about 86 in each bin so, we used half the size of the bin for δ . Then we shift one domain by a width of time-delay constant δ_{\max} which is the lag with maximum correlation value.

The plots below in Figure 21 is called correlogram, x-axis gives the lag and y-axis gives the correlation, r_{δ} at each lag represented by vertical bars in the plot (Metcalfé et al. (2009)). The horizontal blue dotted line indicates confidence interval (CI), which is set to 90%.

We checked if all window pairs are within proximity or not and found out that all anomaly pairs were within the defined proximity.

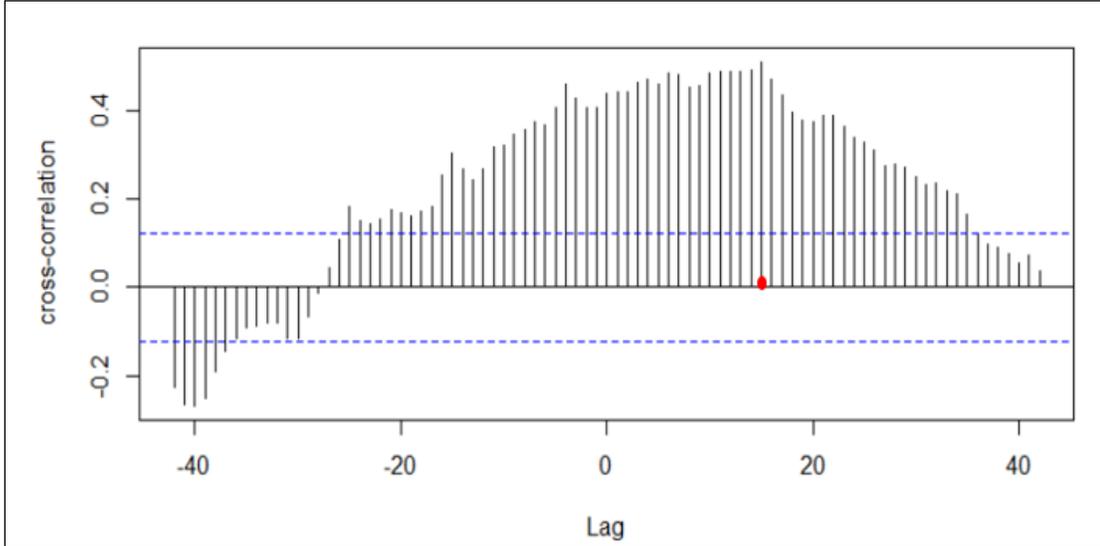


Figure 21: Correlogram between Temperature and Humidity in Bin 2

From the Figure 21, we can see that highest correlation value is at lag 15 which is our δ_{\max} . Hence, we shift one domain with time event width of $15t$ and discover the relation using our novel association algorithm. The discovered relations are listed in Table 20.

Relation Found	Support	Confidence	Lift
$\{t2_radiation\} \Rightarrow \{t2_humidity\}$	0.3571	0.6667	1.4358
$\{t2_humidity\} \Rightarrow \{t2_radiation\}$	0.3571	0.7692	1.4358
$\{t2_humidity\} \Rightarrow \{t2_temperature\}$	0.3214	0.6923	1.0769
$\{t2_temperature\} \Rightarrow \{t2_humidity\}$	0.3214	0.5000	1.0769
$\{t2_radiation, t2_temperature\} \Rightarrow \{t2_humidity\}$	0.2142	1.0	2.1538

{t2_humidity, t2_temperature} => {t2_radiation}	0.2142	0.6667	1.2444
--	--------	--------	--------

Table 20: Delayed relation found in weather data

After shifting the surface domain by $\delta_{\max} = 15t$, we found relations between all three domains. In the third and fourth relations, we have a relation between humidity and temperature with decent support, confidence and lift values. In the fifth relation, we have a relation between all three domains with high confidence and lift values. This indicates that Humidity domain has delayed relation with the Temperature and Solar Radiation domains.

4.3.5 Performance Comparison and Accuracy Evaluation

In this section, we present the various means of accuracy and performance evaluation we conducted to analyze the output of our anomaly detection technique. We employ the Contingency table, Correlation and PAA to validate the anomalies picked up by the scan statistic.

4.3.5.1 Contingency Table

In this section, we present the result we achieved after analyzing anomalous windows that were discovered by scan statistic in synthetic data and anomalous windows we inserted in the synthetic data. Our goal here is to measure the performance of the scan statistic using contingency table which evaluates how many of anomalous time events inserted by us were picked up by the scan statistic while analyzing synthetic data. Time events that are in inserted anomalous time windows and got picked up by scan statistic in synthetic data are our True Positives, time events that were missed are False

Negatives, time events that were picked up in synthetic data but were not in the anomalous time windows inserted by us are our False Positives and time events that are not anomaly and didn't get picked up by scan statistic are our True negative. Table 21, Table 23 and Table 25 lists the inserted anomalous windows and anomalous windows discovered in synthetic data for Total Injured, Light Condition and Surface Condition domain respectively and Table 22, Table 24 and Table 26 presents the contingency table for the Total Injured, Light Condition and Surface Condition domain respectively. Using the contingency table, we calculated the Precision, Recall, and Accuracy which are listed in Table 27.

Bin No.	Inserted anomalies	Synthetic data
1	2014/6/1 – 2014/6/6	2014/6/1 – 2014/6/9
2	2014/6/13 – 2014/6/17	2014/6/13 – 2014/6/18
3	2014/6/22 – 2014/6/27	2014/6/19 – 2014/6/27

Table 21: Anomalies in real-world and synthetic data of Total Injured

	Discovered	
Inserted & real-world anomalies	True	False
True	17	0
False	7	6

Table 22: Contingency table for Total Injured

Bin No.	Inserted anomalies	Synthetic data
1	2014/6/1 – 2014/6/6	2014/6/4 – 2014/6/9
2	2014/6/13 – 2014/6/17	2014/6/13 – 2014/6/18
3	2014/6/22 – 2014/6/27	2014/6/22 – 2014/6/27

Table 23: Anomalies in real-world and synthetic data of Light Condition

	Discovered	
Inserted & real-world anomalies	True	False
True	14	3
False	4	9

Table 24: Contingency table for Light Condition

Bin No.	Inserted anomalies	Synthetic data
1	2014/6/1 – 2014/6/6	2014/6/1 – 2014/6/6
2	2014/6/13 – 2014/6/17	2014/6/13 – 2014/6/18
3	2014/6/22 – 2014/6/27	2014/6/22 – 2014/6/27

Table 25: Anomalies in real-world and synthetic data of Surface Condition

	Discovered	
Inserted & real-world anomalies	True	False
True	17	0
False	1	12

Table 26: Contingency table for Surface Condition

Domain	Precision	Recall	Accuracy
Total Injured	0.7	1	0.76
Light Condition	0.78	0.82	0.76
Surface Condition	0.94	1	0.967

Table 27: Precision, Recall and Accuracy

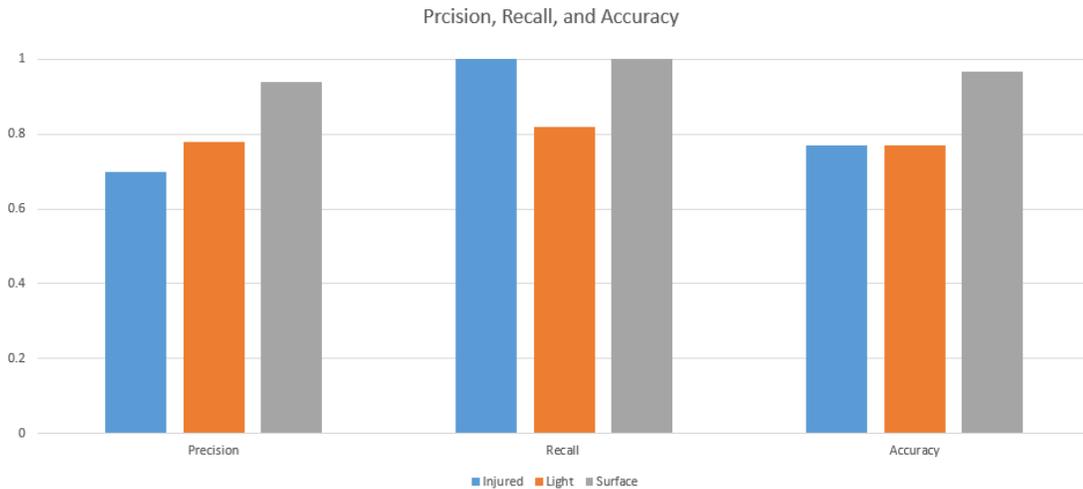


Figure 22: Precision, Recall and Accuracy plot

Looking at Table 27 and Figure 22 above, we can see that we achieved a good result for Surface Condition domain. However, for Total Injured and Light Condition, we see relatively lower values for Precision and Accuracy. Recall for all domains are really high because there were less False Negatives, which indicate that scan statistic was able to capture all anomalous time units. Lower values of Precision and Accuracy indicates that scan statistic captured some False Positive, which could be because of existing anomalous time units in real-world data.

4.3.5.2 Correlation

For a set of domains with more overlaps between their anomalous windows pairs, after discovering relations we compute support, confidence and lift as a means of accuracy measures. On top of that, as an added layer of validation, we also compute the correlation between the anomaly pairs. For our NJDOT and synthetic data, we had overlaps in more than 50 % of anomalous window pairs which implies a stronger direct relation between domains in that dataset. So, we computed correlation as another performance measure.

4.3.5.2.1 Correlation in NJDOT data

We use anomalous time window pairs in each bin and correlate them using readily found correlation technique. Pearson correlation values r for every pair of domains in each bin is listed in Table 28.

Domain pairs	Bin No.	Correlation (r)	Relation
Total Injured and Light Condition	Bin 1	0.049	✓
Total Injured and Light Condition	Bin 2	-0.023	✓
Total Injured and Light Condition	Bin 3	0.051	✓
Light Condition and Surface Condition	Bin 1	0.074	✗
Light Condition and Surface Condition	Bin 2	-0.094	✓
Light Condition and Surface Condition	Bin 3	0.013	✓
Total Injured and Surface Condition	Bin 1	-0.123	✗
Total Injured and Surface Condition	Bin 2	0.055	✓

Total Injured and Surface Condition	Bin 3	0.088	✓
-------------------------------------	-------	-------	---

Table 28: Correlation for all domain pairs in each bin (NJDOT)

The absolute value of correlation higher than zero indicates some correlation. In Table 28, we can see that every pair of domains has values higher than zero which indicates that there some relation between domains, Total Injured, Light Condition, and Surface Condition.

4.3.5.2.2 Correlation in synthetic data:

We use anomalous time windows pair in each bin and correlate them using readily found Pearson correlation technique. The correlation found in each bin between Total Injured, Light Condition and Surface Condition are shown in Table 29.

Domain pairs	Bin No.	Correlation (r)	Relation
Total Injured and Light Condition	Bin 1	0.027	✓
Total Injured and Light Condition	Bin 2	0.000	✓
Total Injured and Light Condition	Bin 3	-0.036	✓
Light Condition and Surface Condition	Bin 1	0.063	✓
Light Condition and Surface Condition	Bin 2	0.067	✓
Light Condition and Surface Condition	Bin 3	0.113	✓
Total Injured and Surface Condition	Bin 1	0.005	✓

Total Injured and Surface Condition	Bin 2	0.139	✓
Total Injured and Surface Condition	Bin 3	-0.020	✓

Table 29: Correlation for all domain pairs in each bin (Synthetic data)

Correlation between two variables indicates that they are not independent and have some impact of one on the other. In Table 29 above, we can see that there are correlation values with an absolute value greater than 0, which indicates that there is some relation between domain Total Injured, Light Condition, and Surface Condition. However, the correlation value for Total Injured and Light Condition in Bin 2 is zero regardless of significant overlaps between these domains in Bin2. This indicates that correlation method is not able to completely capture the existing significant relation between domains that was discovered by our approach because our approach captures relation based on co-occurrences while correlation is based on values of data points.

4.3.5.3 PAA

In this section, we have listed Figures we achieved after plotting the values for each domain. For PAA, we have used the width of $3t$ for each section. The horizontal red bold line indicates PAA value which is the mean of $3t$.

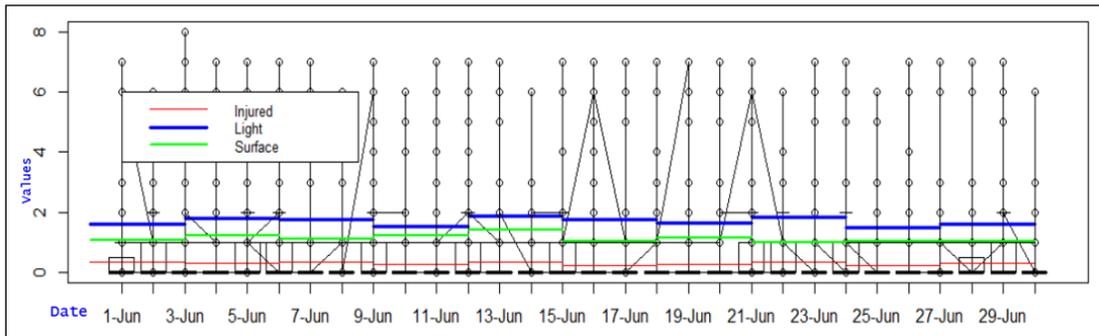


Figure 23: PAA plot for all domains (NJDOT)

In Figure 23, we merged the PAA values for all domains in one plot for NJDOT data. For Total Injured domain, we can see data points with higher values in time window 2014/6/1 – 2014/6/9 which is anomalous time window discovered by scan statistic listed in Table 14, but the mean values that were produced by PAA aren't very significant. For time windows, 2014/6/13 – 2014/6/15 and 2014/6/19 – 2014/6/24 also we can only see slightly higher mean values. In the Light Condition, we can see slightly higher values for all three anomalous time windows: 2014/6/4 – 2014/6/9, 2014/6/13 – 2014/6/18 and 2014/6/19 – 2014/6/24. In the Surface Condition domain, we can see that mean values for time-windows 2014/6/10 – 2014/6/12 and 2014/6/13 – 2014/6/15 are higher than others and mean value for time window 2014/6/19 – 2014/6/21 is only slightly higher.

We can see that overlaps between anomalous time windows for all domains are not quite clearly visible because of very slight variation of mean values.

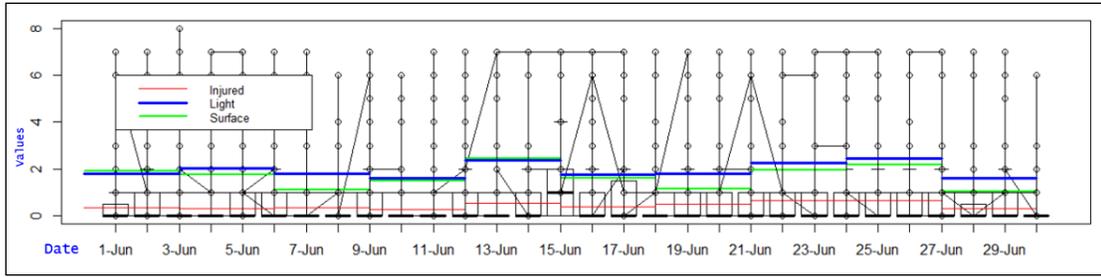


Figure 24: PAA plots for all domains (Synthetic data)

In Figure 24, we merged the PAA values for all domains in one plot for synthetic. In the Total Injured domain, discovered anomalous time windows are: 2014/6/1 – 2014/6/9, 2014/6/13 – 2014/6/18, and 2014/6/19 – 2014/6/27. We can see that mean values for all anomalous time windows are higher than normal time windows i.e. 2014/6/10 – 2014/6/12 and 2014/6/28 – 2014/6/30. In the Light Condition domain, discovered anomalous time windows are: 2014/6/4 – 2014/6/9, 2014/6/13 – 2014/6/18, and 2014/6/22 – 2014/6/27 and we can see that mean values for all anomalous time windows are higher than normal time windows. In the Surface Condition domain, discovered anomalous time windows are: 2014/6/1 – 2014/6/6, 2014/6/13 – 2014/6/18, and 2014/6/22 – 2014/6/27 and we can see that mean values for all anomalous time windows are higher than normal time windows. Compared to Figure 23, we can see anomalous time windows overlaps much clearly here in Figure 24.

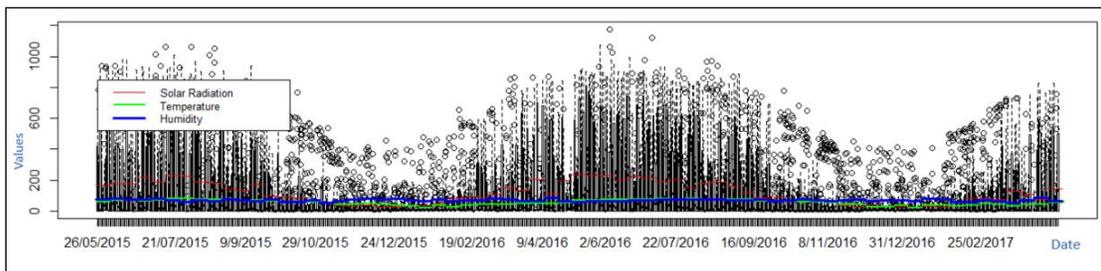


Figure 25: PAA plot for Solar Radiation, Temperature, and Humidity (Weather)

In Figure 25, we merged the PAA plots for all three domains in weather data: Solar Radiation, Temperature, and Humidity. The anomalous time windows aren't very clearly visible in the plot but we can see the trend going on in the data for all three domains. We can see that when values of Solar Radiation and Temperature is higher then, values in Humidity is lower which indicated that Humidity has delayed relation with Solar Radiation and Temperature domain.

Analyzing the PAA values in plots above, we can see that PAA was able to capture most of the anomalous time windows in synthetic data very well but not in NJDOT data. The result is more visible in synthetic data than in real-world data because of the subtle transition of mean values of an anomalous time window in real-world data while the transition is quite significant in synthetic data because we inserted unusually higher data points. So, we can infer that PAA method is capable of identifying unusual series of events if there is a significantly high number of unusual events but may not perform well for few unusual events in a smaller time sequence. However, scan statistic is capable of identifying both, unusual series of events with a significantly high number of unusual events and few unusual events in a smaller time sequence which indicates that scan statistic is better at identifying unusual events.

4.3.5.4 Comparison with larger dataset

For NJDOT data, we used one month of data for our analysis. We wanted to see if scan statistic could discover those anomalies while using larger data or not. So, we used the whole year of data for the year 2014 for Injured and Surface Condition domain to get

anomalies using scan statistic. Anomalies that were discovered is presented in the Table 30 and shown in Figure 26 below.

Total Injured	Surface Condition
2014/1/15 to 2014/2/1	2014/2/2 to 2014/2/16
2014/4/15 to 2014/5/2	2014/2/20 to 2014/2/22
2014/5/30 to 2014/6/10	2014/5/15 to 2014/5/17
2014/8/1 to 2014/8/3	2014/7/2 to 2014/7/4
2014/10/3 to 2014/10/5	2014/10/21 to 2014/10/23
2014/12/14 to 2014/12/28	2014/11/26 to 2014/12/10

Table 30: Anomalies discovered in one year data (NJDOT)

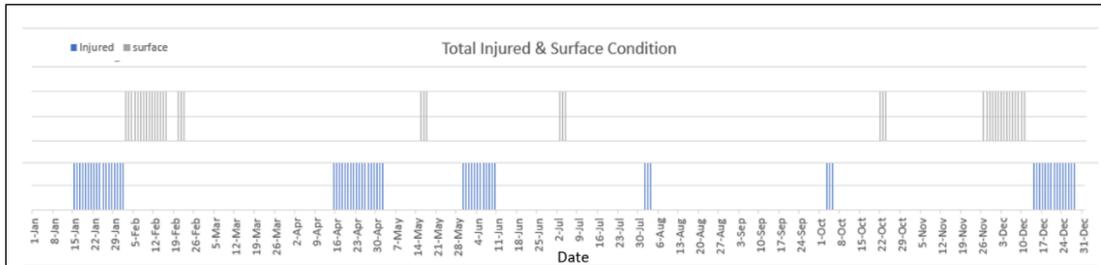


Figure 26: Anomalies in whole year data (NJDOT)

In Table 30, we can see that we have anomaly discovered in the Total Injured domain for the month of June, which was also discovered in our one-month data listed in Table 14, which indicates that this was a significant anomalous time window. However, if we look into the discovered anomaly for Surface Condition domain, we don't have any anomalies for the month of June even though we were able to find few in our one-month data. If we look at Figure 26, we can see that there are significant anomalous time windows during the month of February and December, which is expected given

that these are months with frequent snowfall and inclement weather. Compared to the surface condition during months of December to March, other months have a decent surface condition, resulting into smaller and weaker anomalies. Also, these weaker anomalies get swamped by the significant anomalies while analyzing larger data.

Chapter 5: Conclusion and Future works

In this thesis, we have proposed a novel algorithm to discover temporal associations across multiple distinct domains using the time windows with unusual events happening. This algorithm allows us to address complex real-world problems by revealing the pattern causing the problem and interrelated domains that have an impact on it. We used three publicly available multi-domain datasets, MATCH, NJDOT and weather data to simulate a real-world problem scenario for our thesis research.

In our multi-domain temporal anomaly association approach, we not only focus on direct impacts on one domain by the other but also to complement the approach we analyze delayed relation between domains as well. We employ the concept of overlap and proximity to discover the direct and the time-delayed relations. We used scan statistic to discover anomalous time windows and validate anomalous windows discovered by scan statistic, we used PAA method. We also measured the accuracy of scan statistic results using contingency table. We used Apriori method to get relations between distinct domains and used correlation to validate relations.

In future, we plan to extend our approach to get relations from n temporal domains with different time-resolution and work on supplementing our approach with good data visualization techniques to present our discovered anomalies and relations. We would also like to integrate the R application we are using with big data framework to be able to analyze voluminous data.

Bibliography

- [1] Antunes, C. M., & Oliveira, A. L. (2001, August). Temporal data mining: An overview. In *KDD workshop on temporal data mining* (Vol. 1, p. 13).
- [2] Nazerfard, E., Rashidi, P., & Cook, D. J. (2011). Using association rule mining to discover temporal relations of daily activities. In *Toward Useful Services for Elderly and People with Disabilities* (pp. 49-56). Springer Berlin Heidelberg.
- [3] Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. *ACM SIGMOD Record*, 22(2), 207-216.
- [4] Liang, Z., Xinming, T., & Wenliang, J. (2005, August). Temporal association rule mining based on T-Apriori algorithm and its typical application. In *Intl. Symposium on Spatial-Temporal Modeling Analysis* (Vol. 5, No. 2).
- [5] Nair, B. B., Mohandas, V. P., Nayanar, N., Teja, E. S. R., Vigneshwari, S., & Teja, K. V. N. S. (2015). A Stock Trading Recommender System Based on Temporal Association Rule Mining. *SAGE Open*, 5(2), 2158244015579941.
- [6] Zhou, H., Wei, W., Shimada, K., Mabu, S., & Hirasawa, K. (2008, June). Time related association rules mining with attributes accumulation mechanism and its application to traffic prediction. In *Evolutionary Computation, 2008. CEC 2008.(IEEE World Congress on Computational Intelligence). IEEE Congress on* (pp. 305-311). IEEE.
- [7] Yamtani, Y., & Ohira, M. (2014, November). An Exploratory Analysis for Studying Software Evolution: Time-Delayed Correlation Analysis. In *2014 6th International Workshop on Empirical Software Engineering in Practice (IWESEP)* (pp. 13-18). IEEE.
- [8] Liang, M., Xi-Hai, L., Wan-Gang, Z., & Dai-Zhi, L. (2015, September). The Generalized Cross-Correlation Method for Time Delay Estimation of Infrasound Signal. In *2015 Fifth International Conference on Instrumentation and Measurement, Computer, Communication and Control (IMCCC)* (pp. 1320-1323). IEEE.

- [9] Harms, Sherri K, Temporal Association Rule Mining in Event Sequence, InEncyclopedia of Data Warehousing and Mining, ed. John Wang, 1098-1102 ,2005
- [10] Roddick, J., Spiliopoulou, M.: A Survey of Temporal Knowledge Discovery Paradigms and Methods. In *IEEE Transactions of Knowledge and Data Engineering*, vol. 13, 2001.
- [11] Bouandas, K., & Osmani, A. (2007, March). Mining association rules in temporal sequences. In *Computational Intelligence and Data Mining, 2007. CIDM 2007. IEEE Symposium on* (pp. 610-615). IEEE.
- [12] Teng, M. (2010, December). Anomaly detection on time series. In *Progress in Informatics and Computing (PIC), 2010 IEEE International Conference on* (Vol. 1, pp. 603-608). IEEE.
- [13] Janeja, V. P., & Palanisamy, R. (2013). Multi-domain anomaly detection in spatial datasets. *Knowledge and information systems*, 36(3), 749-788.
- [14] Golmohammadi, K., & Zaiane, O. R. (2015, October). Time series contextual anomaly detection for detecting market manipulation in stock market. In *Data Science and Advanced Analytics (DSAA), 2015. 36678 2015. IEEE International Conference on* (pp. 1-10). IEEE.
- [15] Chan, P. K., & Mahoney, M. V. (2005, November). Modeling multiple time series for anomaly detection. In *Data Mining, Fifth IEEE International Conference on* (pp. 8-pp). IEEE.
- [16] Harms, S. K., Saquer, J., Deogun, J., & Tadesse, T. (2001). Discovering Representative Episodal Association Rules from Event Sequences Using Frequent Closed Episode Sets and Event Constraints, *Proc. ICDM '01* (pp. 603-606). Silicon Valley, CA.
- [17] Ramaswamy, S., Mahajan, S., & Silberschatz, A. (1998, August). On the discovery of interesting patterns in association rules. In *VLDB* (Vol. 98, pp. 368-379).
- [18] Rebbapragada, U., Protopapas, P., Brodley, C. E., & Alcock, C. (2009). Finding anomalous periodic time series. *Machine learning*, 74(3), 281-313.

- [19] Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics-Theory and methods*, 26(6), 1481-1496.
- [20] Faires, M. C., Pearl, D. L., Ciccotelli, W. A., Berke, O., Reid-Smith, R. J., & Weese, J. S. (2014). The use of the temporal scan statistic to detect methicillin-resistant *Staphylococcus aureus* clusters in a community hospital. *BMC infectious diseases*, 14(1), 1.
- [21] Kulldorff, M. (2001). Prospective time periodic geographical disease surveillance using a scan statistic. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 164(1), 61-72.
- [22] Taylor, M. (2015). Sinkr: sinkr. R package version 2.0. Retrieved March 24, 2017, from <https://github.com/marchtaylor/sinkr>
- [23] Haight, F. A. (1967). *Handbook of the Poisson distribution* (No. 519.23 H3).
- [24] Unit, N. W. (n.d.). Crash Records. Retrieved March 24, 2017, from <http://www.state.nj.us/transportation/refdata/accident/>
- [25] Rankings Data. (n.d.). Retrieved March 24, 2017, from <http://www.countyhealthrankings.org/rankings/data>
- [26] Gorman, R. M. (2009). Intercomparison of methods for the temporal interpolation of synoptic wind fields. *Journal of Atmospheric and Oceanic Technology*, 26(4), 828-837.
- [27] Zavala-Hidalgo, J., Bourassa, M. A., Morey, S. L., O'Brien, J. J., & Yu, P. (2003, September). A new temporal interpolation method for high-frequency vector wind fields. In *OCEANS 2003. Proceedings* (Vol. 2, pp. 1050-1053). IEEE.
- [28] Buuren, S., & Groothuis-Oudshoorn, K. (2011). [22: Multivariate imputation by chained equations in R. *Journal of statistical software*, 45(3).
- [29] Lin, J., Keogh, E., Lonardi, S., & Chiu, B. (2003, June). A symbolic representation of time series, with implications for streaming algorithms. In *Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery* (pp. 2-11). ACM.
- [30] Metcalfe, A. V., & Cowpertwait, P. S. (2009). *Introductory time series with R*.

[31] Beach Weather Stations - Automated Sensors | City of Chicago | Data Portal.
(n.d.). Retrieved April 25, 2017, from <https://data.cityofchicago.org/Parks-Recreation/Beach-Weather-Stations-Automated-Sensors/k7hf-8y75>

