Sequential Spatio-Temporal Pattern Mining with Time Lag

BY

PAVAN REDDY
B.E., Visvesvaraya Technological University, India, 2007

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Computer Science
in the Graduate College of the
University of Illinois at Chicago, 2014

Chicago, Illinois

Defense Committee:

Isabel Cruz, Chair and Advisor
Stanley Sclove, Information & Decision Sciences
Brian Ziebart

To my parents,

thanks for your love and support

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my advisor Prof. Isabel F. Cruz for the continuous support, patience, motivation, and immense knowledge. Her guidance helped me in my research and writing of this thesis.

I would like to thank the rest of my thesis committee: Prof. Stanley L. Sclove, and Prof. Brian Ziebart for their encouragement and insightful comments.

I want to thank Claudio Caletti for his work on spatial resolution.

I thank everyone at the Advances in Data, Visual, and Information Science Research Laboratory (ADVIS); especially Venkat R. Ganesh.

Finally, I thank my sister for her help during my stay in the United States.

PR

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

GIS	Geographic Information Systems.
MOWCATL	Minimal Occurrences With Constraints and Time Lags.
MDCOP	Mixed Drove spatio-temporal Co-occurrence Patterns.
GIVA	Geospatial and Temporal Data Integration, Visualization, and Analytics.
RMSE	Root Mean Squared Error.
ANOVA	Analysis Of Variance.
DF	Degrees of Freedom.
SS	Sum of Squares.
MS	Mean Squares.
NASA	National Aeronautics and Space Administration.
TRMM	Tropical Rainfall Measuring Mission.
USGS	United States Geological Survey.

SUMMARY

Geographic Information Systems have a variety of applications. There are many large spatio-temporal datasets available. Performing spatial and temporal data mining and analysis on these datasets enable knowledge discovery, which help domain experts discover unknown and interesting spatial and temporal insights in the data. This thesis investigates spatial and temporal relationships between two different spatial datasets and proposes a method to mine spatio-temporal patterns from the spatial datasets. Machine learning techniques are used to make predictions using historical data based on the spatio-temporal patterns.

CHAPTER 1

INTRODUCTION

An episode is a sequence of events occurring at different locations at different times, that is occurrence of events at the same or different locations separated by a delay in time. For example, an event e_1 occurs at location A at time t_1 and is succeeded by another event e_2 at location B at time t_2 where the events are separated by a time delay of (t_2 - t_1). Spatial association rules and patterns are defined as frequently occurring episodes in spatial datasets. Spatial patterns can be spatial association rules, co-location patterns and spatio-temporal co-occurrence patterns.

Figure 1 displays a representation of events and episodes on a time line. Events are occurrences in the dataset, for example, rainfall of r_1 mm at location loc_1 at time t_2. Episode

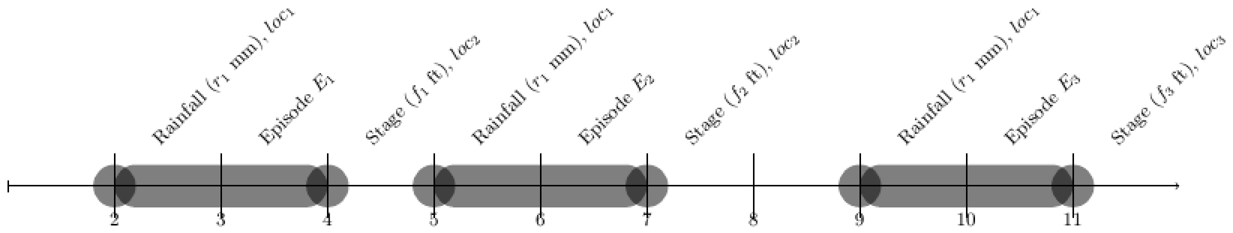
Rainfall (r_1 mm), loc_1
Episode E_1
Stage (f_1 ft), loc_2
Rainfall (r_1 mm), loc_1
Episode E_2
Stage (f_2 ft), loc_2
Rainfall (r_1 mm), loc_1
Episode E_3
Stage (f_3 ft), loc_3
2
3
4
5
6
7
8
9
10
11

Figure 1. Representation of events and episodes on a time line

is a sequence of events, for example, E_1 is an episode of rainfall occurring at t_2 succeeded by increase in the surface water stage at t_4. Spatial features are discrete geographic objects at a particular location, such as a library in a spatial dataset of a city. Boolean spatial features describe the presence or absence of a geographic object at a location. Co-location patterns (1) are patterns of boolean spatial features that occur frequently in proximity of other spatial features, such as a library in close spatial proximity of a school frequently occurring in a dataset of a city, where the library and school are spatial features. Mixed Drove spatio-temporal Co-occurrence Patterns (MDCOPs) (2) are similar to co-location patterns but they also include the temporal influence between the features. Discovering MDCOPs is a technique where the dataset is mined for spatio-temporal features that are in close proximity to other features on both the spatial and temporal scales.

Supervised learning (3) is a machine learning technique of inferring a relationship between different variables using labeled training data. Training data is a set of examples that are used to build a machine learning model. Regression (3) can be used as a supervised machine learning technique that estimates the behavior of a dependent variable based on one or more independent variables. Regression analysis can be used to analyze historical data and infer a relationship between the dependent and independent variables, and use this inferred relationship to make predictions about the behavior of the dependent variable in the future. Predictive analysis is a data analysis technique of using historical facts to make predictions about the future. Regression can be used for predictive analysis.

In this thesis, we use hydro-metric spatial datasets of precipitation and surface water-gage for mining spatio-temporal patterns of rainfall and surface water. Stage and flow are measures of surface elevation and volumetric discharge of surface water. Spatial datasets are mined for spatio-temporal patterns of rainfall and surface water stage and these patterns are used to build regression models that estimate surface water stage based on rainfall data.

1.1 Problem Definition

Given two spatial datasets, the motivation is to establish a relationship between the datasets and investigate if the datasets are correlated with or without a delay in time. If the datasets are correlated then the time delay between the datasets is estimated. The datasets are mined for spatio-temporal patterns that occur frequently in the data. Based on these spatio-temporal pattens, regression models should be built that can be used to forecast the dependent variable behavior.

CHAPTER 2

RELATED WORK

Discovering spatio-temporal patterns is a very important problem and there have been various methods and approaches used in this field of research.

2.1 Spatio-temporal pattern mining

2.1.1 Minimal Occurrences With Constraints and Time Lags

Minimal Occurrences With Constraints and Time Lags (MOWCATL) (4) is an efficient algorithm for mining frequent association rules in a dataset. Time lag is taken into account in the association rules between occurrence of events in an episode. This approach uses a sliding window on the time scale of a predefined size. The dataset is also converted to the finest granularity on the time scale before discovering association rules. The dataset is discretized by using a clustering algorithm to label clusters of events. Then given a target episode and a time lag, patterns are discovered that contain episodes of a given maximum window width and a maximum time lag.

2.1.2 Co-location patterns

Co-location patterns (1) are patterns of boolean spatial features/objects that are located frequently in spatial neighborhood of other spatial features/objects. Co-location patterns use a threshold distance value between spatial objects. Co-location patterns are mined using either a spatial statistical method or an association rule mining method. Spatial statistical method makes use of Ripley's K function and statistical correlation to identify spatial features in proximity of other spatial features. Association rule mining method identifies occurrence of events in a dataset and uses data mining algorithms such as apriori algorithm to discover co-location patterns.

2.1.3 Mixed Drove spatio-temporal Co-occurrence Patterns

Mixed Drove spatio-temporal Co-occurrence Patterns (MDCOPs) (2) are spatio-temporal patterns of spatial features that occur in proximity of other spatial features on both space and time scale.

2.1.4 Interval orientation patterns

Interval orientation patterns (5) are patterns that consider the duration of the existence of spatial features. This approach enables discovering patterns that involve spatial features which occur for a finite duration of time at a particular location. This method is useful for finding patterns of activity detection in spatial and temporal data.

2.2 Relationship between precipitation and shallow ground water

Relationship between precipitation and shallow ground water (6) has been studied by using statistical time series analysis to establish a relationship between monthly precipitation and groundwater (water table) levels. This is used to investigate physical relationships between precipitation, groundwater, geophysical conditions (topography, soil etc.) and drought analysis. Similarly in this thesis, we establish a relationship between rainfall and surface water.

CHAPTER 3

METHOD AND ALGORITHM

3.1 Data

Spatial and temporal datasets of rainfall and surface water is used in this thesis. The rainfall data, measured in millimeters (mm), is obtained from the NASA Tropical Rainfall Measuring Mission (TRMM)[1] and the surface water-gage data is obtained from the United States Geological Survey (USGS)[2].

The surface water data is collection at water-gage stations. Various attributes of surface water are collected by the surface water-gages such as flow and stage. Flow or volumetric discharge is the volume of water passing per second through a section of the water-gage station. Flow is measured in cubic feet per second (cf/s). Stage or gage height is the elevation of the water surface above the vertical datum. The vertical datum is a predefined 'zero' point that is constant for a station. Since the vertical datum is different for each station, the stage values are not comparable for different water-gage stations. Stage is measured in feet (ft). Flood stage is the elevation at which the water overflows from the natural banks of the stream. Stage margin is the difference between the stage and the flood stage of a station. Both, flood stage and stage margin are measured in feet (ft).

---

[1] http://trmm.gsfc.nasa.gov/

[2] http://www.usgs.gov/

Precipitation and surface water data used in this thesis are from the state of Illinois of the United States of America.

3.2 Spatial Resolution

Based on the data acquisition method, heterogeneities are introduced in the datasets. Rainfall data is recorded by NASA TRMM at specific locations on a 0.25 x 0.25 mile grid and surface water data is recorded at water-gage stations.

This heterogeneity makes it necessary for data from different sources to be normalized to a common spatial resolution. Spatial autocorrelation (7) is used to resolve the heterogeneity in the spatial resolution of the two datasets (8). The data is normalized to rectangular tessellations to form a grid as shown in Figure 2. Each cell in the grid acquires a value of the weighted average of the data points within the cell. Spatial autocorrelation (9) is characterized by correlation in the neighboring locations in space. Moran's I index, ranging from -1 to +1, is the measure used for spatial autocorrelation. A positive Moran's I value indicates clustering of similar data points. A negative Moran's I value indicates dissimilarity of the value of a data point with its neighbors and a Moran's I value that is close to zero indicates randomness in the data. The dimensions of a grid with rectangular tessellations is computed based on the value at which the Moran's I index is maximum. This method (9) was the work of Claudio Caletti.

In this thesis, the grid cells are defined by their row and column indices e.g., cell (2,3) represents the cell in the 2^{nd} row and 3^{rd} column of the grid. A cell can also be represented by latitude

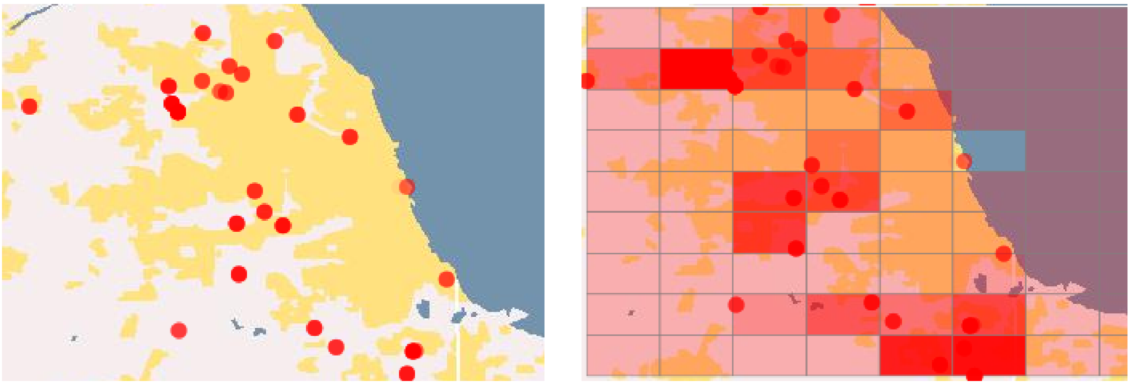

Figure 2. Spatial resolution - Raw data points and tessellations of optimal spatial resolution

and longitude of the points defining its top-left and bottom-right co-ordinates e.g., (41.323 N, -88.073 W), (40.597 N, -87.5 W).

3.3 Temporal Resolution

Data from different sources can be acquired at different frequencies of data collection, for example, rainfall data is collected once a day whereas surface water data is collected twice a day. It is necessary to normalize datasets of different temporal resolutions to a common temporal resolution. The common temporal resolution is chosen to be equal to the resolution of the dataset with the least frequent data collection method. The dataset with the higher frequency data collection method is aggregated by computing the mean of the data points. For example, if dataset D_1 is recorded daily and dataset D_2 is recorded twice a day, then the data is normalized to the resolution of dataset D_1, and the values for dataset D_2 are aggregated by computing the mean of data points aggregated by a day.

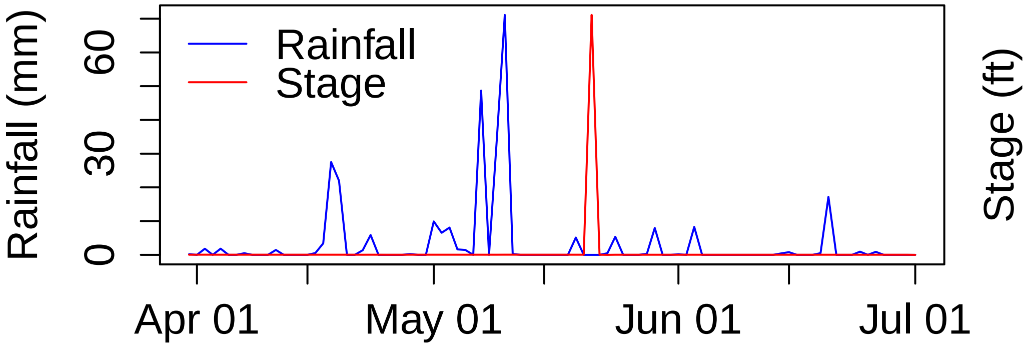

Figure 3. Rainfall (mm) and Surface water stage (ft) for grid cells with co-ordinates (41.323 N, -88.073 W), (40.597 N, -87.5 W) and (36.966 N, -88.647 W), (37.692 N, -88.073 W) on the respective grids

3.4 Time Delay

Statistical cross correlation (10) is used to estimate the time delay between two time series. Cross correlation is employed to find if rainfall and water-gage data is correlated with each other with or without a time delay. Cross correlation is a measure of similarity of two sets of time series data while a time delay is applied to one of them.

$$r_{xy}[\tau] = \frac{\sum_t (x[t] - \mu_x])(y[t-\tau] - \mu_y)}{\sqrt{\sum_t (x[t] - \mu_x])^2}\sqrt{\sum_t (y[t-\tau] - \mu_y)^2}} \tag{3.1}$$

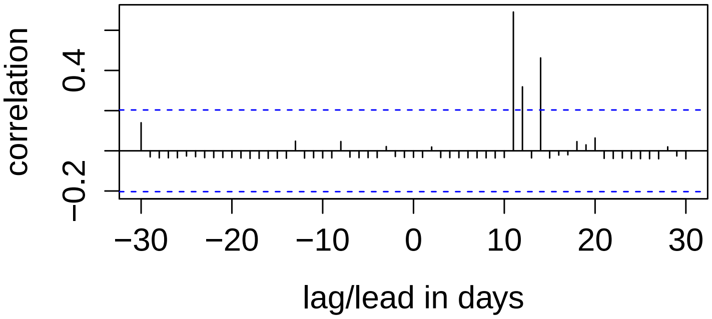

Figure 4. Correlation of Rainfall (mm) and surface water stage (ft) having a maximum correlation of 0.691 after a time delay of 11 days

In Equation 3.1, $r_{xy}[\tau]$ is the cross correlation of x[t] and y[t], x[t] and y[t] are the time series and τ is the time delay. μ_x and μ_y are the mean values of the time series x[t] and y[t]. Cross correlation is computed for different values of τ and the delay with the maximum correlation is used to compute the actual time delay between the two time series, as shown in Equation 3.2.

$$\hat{\tau} = argmax_{\tau}(r_{xy}[\tau]) \tag{3.2}$$

In Equation 3.2, $\hat{\tau}$ represents the time delay where the maximum correlation is achieved and $r_{xy}[\tau]$ is the cross correlation of x[t] and y[t] computed at time delay τ.

The time delay for which the maximum value of correlation is obtained using cross correlation signifies that the datasets tend to be similar at this time delay.

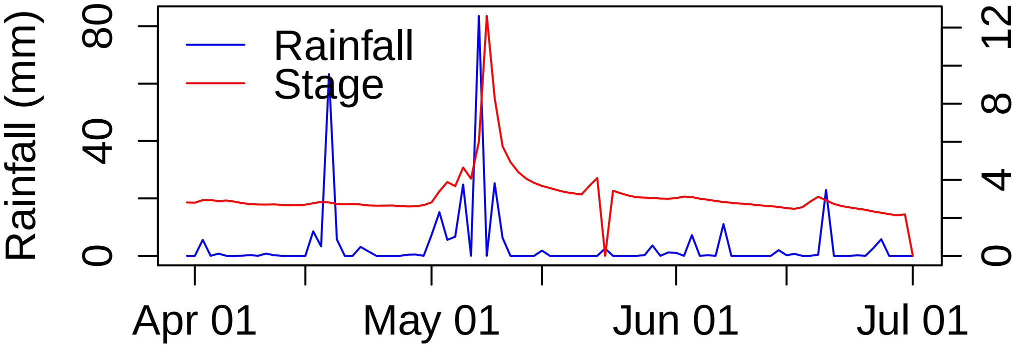

Figure 5. Rainfall (mm) and surface water stage (ft) for grid cells with co-ordinates (42.05 N, -89.794 W), (41.323 N, -89.221 W) and (40.597 N, -88.647 W), (41.323 N, -88.073 W) on the respective grids

Figure 3 and Figure 4 show the correlation between rainfall and surface water stage is 0.691 with a time delay of 11 days at tessellation with top-left and bottom-right co-ordinates of the cell (41.323 N, -88.073 W), (40.597 N, -87.5 W) and tessellation with top-left and bottom-right co-ordinates (36.966 N, -88.647 W), (37.692 N, -88.073 W) on the respective grids. Figure 5 and Figure 6 display the relationship between rainfall and stage, the delay computed between rainfall and stage is 1 day at the tessellation with top-left and bottom-right co-ordinates of the cell (42.05 N, -89.794 W), (41.323 N, -89.221 W) and tessellation with top-left and bottom-right co-ordinates (40.597 N, -88.647 W), (41.323 N, -88.073 W) on rainfall and surface water grids.

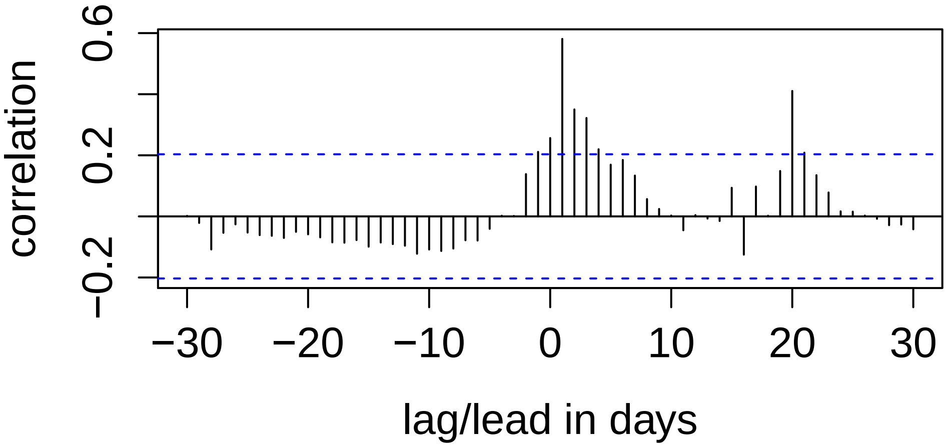

Figure 6. Correlation of Rainfall (mm) and surface water stage (ft) having a maximum correlation after a time delay of 1 day

3.5 Spatio-temporal pattern mining

A variant of the apriori algorithm (11) is used to discover frequently occurring spatio-temporal patterns in the datasets. The apriori algorithm consists of two stages: candidate generation and pattern pruning. In the candidate generation stage, candidate patterns are generated for each pair of tessellations of both rainfall and water-gage grids. For each pair of tessellations of the grids of both datasets, the time delay is computed using cross correlation. The candidate patterns are then subject to pattern pruning. In the pattern pruning stage, the support and confidence of each spatio-temporal pattern is computed. Candidates are pruned based on the minimum support and correlation coefficients. Support and confidence measures are discussed in Section 5.1.

3.6 Regression

Regression (3) is used to model a relationship between an independent variables and one or more dependent variables in a dataset. The surface water stage is used as the dependent variable and rainfall, time delay and the locations are used as the independent variables.

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \cdots + \beta_p x_{tp} + \varepsilon_t \quad t = 1, \ldots, n \tag{3.3}$$

In Equation 3.3, p is the number of variables. An example for x is rainfall. The β values are the learned parameters and ε is a disturbance term or error variable. This can be rewritten as Equation 3.4:

$$y_t = \mathbf{x}_t^T \beta + \varepsilon_t, \quad t = 1, \ldots, n \tag{3.4}$$

where $\mathbf{x}_t^T$ represents the transpose of the vector of independent variables and $\mathbf{x}_t^T \beta$ is the inner product between vectors $\mathbf{x}_t^T$ and β. Using the spatio-temporal patterns and the time delay, regression can be modeled to accommodate time delay on the dependent variables as shown in Equation 3.5, that is, rainfall data can be used to build a regression model and by making variations for time delay as show in Equation 3.5, the surface water stage variable can be forecast.

$$y_{t+\tau} = \mathbf{x}_t^T \beta + \varepsilon_t, \quad t = 1, \ldots, n \tag{3.5}$$

Regression algorithms such as linear regression and support vector regression (3) are used for building regression models based on the spatio-temporal patterns. Support vector regression models are further tuned to obtain the best performance by employing techniques such as a kernel method. Kernel methods (3) enable the use of higher dimensional feature spaces.

CHAPTER 4

IMPLEMENTATION

Spatial datasets for both surface water and rainfall are obtained in flat files from the NASA Tropical Rainfall Measuring Mission (TRMM) and the United States Geological Survey (USGS). This data is transformed and imported to a spatial relational database. PostgreSQL[1] is the relational database used in this thesis. PostgreSQL is used with PostGIS[2], a spatial database extender for PostgreSQL. R (12) and Java are used for implementation of the data processing, statistical and machine learning components. Spatial autocorrelation and Moran's I index computation modules are used from GIVA and the ape package (13) for R. The e1071 package (14) for R is used for support vector regression.

---

[1] http://www.postgresql.org/

[2] http://postgis.net/

CHAPTER 5

EVALUATION

5.1 Support and confidence

Spatio-temporal patterns are evaluated using support and confidence values. Support (15) is a measure of how frequently an event or a pattern occurs in the dataset. Confidence (15) is a measure of the predictability of the pattern in the dataset, it can be defined as shown in equation (6), where $rain_{l1}$ represents rainfall at location *l1*, $stage_{l2}$ represents the surface water stage at location *l2* and $(rain_{l1} \rightarrow stage_{l2})$ represents a spatio-temporal pattern.

$$conf(rain_{l1} \rightarrow stage_{l2}) = \frac{supp(rain_{l1} \rightarrow stage_{l2})}{supp(rain_{l1})} \tag{5.1}$$

For example, lets consider the pattern $(rain_{l1} \rightarrow stage_{l2})$. If 1% of all patterns in the dataset are $(rain_{l1} \rightarrow stage_{l2})$ then the support of $(rain_{l1} \rightarrow stage_{l2})$ is 1%, that is equal to 0.01. If 3% of all the rainfall occurs at location l1 then support of $(rain_{l1})$ is 3%, that is equal to 0.03.

If it rains at location l1 then the confidence of the pattern $(rain_{l1} \rightarrow stage_{l2})$ is 0.01/0.03, that is equal to 0.333.

5.2 Root Mean Squared Error

Regression models are used to forecast the dependent variable based on the independent variables. The regression model and predictions are evaluated using the Root Mean Squared Error (RMSE) (3).

The Root Mean Squared Error (RMSE) is a measure of the square root of the mean of the squared error between the predicted and actual values. A lower RMSE value indicates a better model. RMSE can also be used as a measure of how well a model fits the data and if the model is overfit or underfit. If the RMSE of the training set is significantly lower than the RMSE of testing set then it indicates a model that is overfit, in which case methods like regularization are used to tune the model. Regularization (3) is a method of reducing overfitting by imposing a penalty for the complexity of the model

The rainfall data is split in 80:20 ratio for the training and testing set. For each spatio-temporal pattern, a regression model is built encompassing its corresponding time lag.

CHAPTER 6

EXPERIMENTAL RESULTS

6.1 Spatio-temporal patterns

Table I displays the spatio-temporal patterns that are discovered. the corresponding correlation, time lag, support and confidence values for rainfall and surface water stage data. The minimum correlation co-efficient is chosen as 75% of the maximum correlation coefficient and minimum support is chosen as 25% of the maximum support value. These values are user defined input parameters to the apriori algorithm (11) and are chosen manually by trial and error such that the patterns discovered provide interesting information and high confidence values while keeping the number of patterns reasonably small (16).

6.2 Regression

Based on the spatio-temporal patterns extracted, linear regression and support vector regression models are built using the data for all the patterns using a 80:20 training to testing set bias. Each pattern shown in Table II is used to build a regression model incorporating the corresponding time delay as described in Section 3.6. These models are used to predict the surface water stage values based on the rainfall data and time lag for each pattern. These model are evaluated using the Root Mean Square Error (RMSE) metric. Both linear regression and support vector regression models are built for all patterns. The Root Mean Squared Error

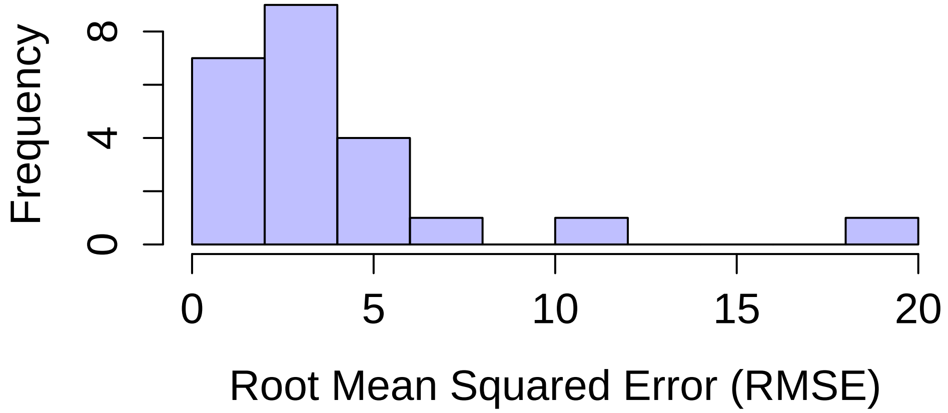

Figure 7. Histogram of the RMSE values for the support vector regression models of all patterns

values for each pattern for both linear and support vector regression is shown in Table II. The support vector regression models shown in Table II use a polynomial kernel of degree 2.

6.2 shows a histogram of the RMSE values for the regression models of all patterns using support vector regression. Low values of the Root Mean Square Error indicate that the regression models perform well.

The watergage grid cell (3,6) with the top-left and the bottom-right co-ordinates (38.418 N, -88.647 W), (39.144 N, -88.073 W) contains one water-gage station, “Little Wabash river below Clay city, IL”. The flood stage of this water-gage station is 16 feet (ft).   6.2 shows

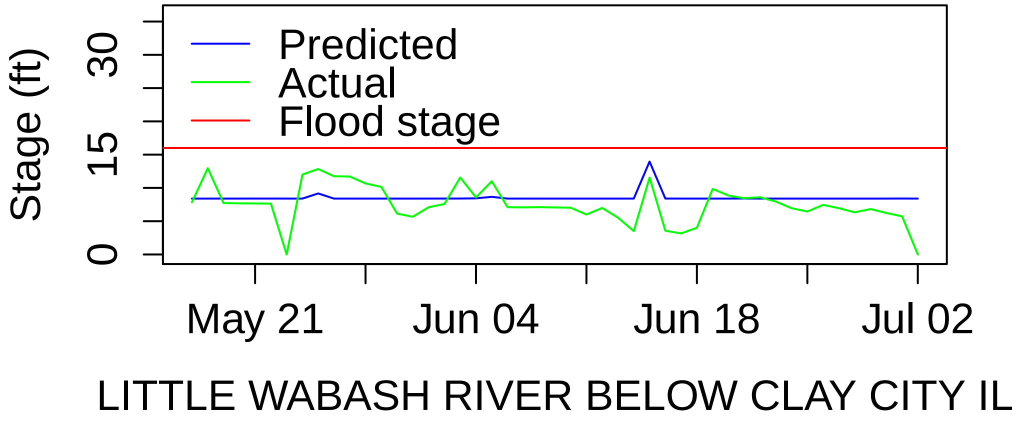

Figure 8. Estimated surface water stage (ft) vs actual surface water stage value, and flood stage margin of the water-gage station "Little Wabash river below clay city IL"

the comparison of the predicted and the actual values of surface water stage (ft) based on the patterns extracted and the corresponding time lag. The figure shows that the predicted estimates of surface water stage are less than the flood stage of the water-gage station. In case the regression model predicts that the surface water stage is greater than the flood stage, then the model estimates a flood at the water-gage station.

6.2.1 Overfitting

As shown in Table I, the Root Mean Square Error (RMSE) values of linear regression for a few patterns is very high for e.g., the pattern $rain_{(2,4)} \rightarrow watergage_{(1,6)}$ has a RMSE value of 3456.16. The grid cell (1,6) with the top-left and bottom-right co-ordinates (36.966,

-88.647),(37.692, -88.073) contains two water-gage stations - “Ohio river at old Shawneetown IL-KY” and “Ohio river at dam 51 at Golconda IL” with flood stages 33 ft and 40 ft respectively. The RMSE for linear regression for the training set is 936.225, which is significantly lower than the RMSE for the testing set, which indicates the model is overfit. The surface water stage data for the watergage cell (1,6) contains a very large outlier which causes the high RMSE values for linear regression because linear regression is sensitive to outliers. To corroborate this finding, the mean of the surface water stage values is computed and it is 464.014. Hence, support vector regression is used and it provides a RMSE value of 11.47. Similarly, high values of RMSE for linear regression are obtained with patterns only involving watergage cell (1,6).

6.2.2 Analysis of Variance

Each cell in the water-gage grid can contain water-gage stations. Based on the flood stage and the estimated stage values of each of these water-gage stations, we can estimate when and if the stage of the water-gage station will be elevated to a value beyond the flood stage and cause flooding of the natural banks of the water-gage station. If there is more than one pattern that affects a grid cell on the water-gage grid, then the average of all the estimates from all regression models is computed. For example, consider the pattern $rain_{(3,2)} \rightarrow watergage_{(3,6)}$ which has a RMSE value 3.123 and 2.77 for linear regression and support vector regression respectively. The Analysis Of Variance (ANOVA) (17) table for linear regression of this pattern is shown in Table III. In Table III, the calculated Sum of Squares (SS) terms are provided in the “Sum of Squares” column, the Mean Square (MS) terms are provided in the “Sum of Squares” column, and the p-value is provided in the "Pr($\leq$F-statistic)" column. "Pr($\leq$F-statistic)" is the probability of observing a value greater than or equal to 17.909.

Similarly, the ANOVA table for linear regression of the pattern $rain_{(6,6)} \rightarrow watergage_{(6,7)}$ is shown in Table IV. A null hypothesis is made that the observed effect in the linear regression model occurs purely by chance. The F-statistic is a statistical test that is used to check if the variances of both the datasets are equal. A high value of F-statistic indicates that the datasets have different variances. A "Pr($\leq$F-statistic)" that is very low indicates that the observed effect is highly unlikely to have occurred purely by chance. Thus, rejecting the null hypothesis.

Grid cell (row,column)		Correlation	Lag	Support	Confidence
Rainfall	Stage				
(6,1)	(1,6)	0.580	14	0.0048	0.469
(2,7)	(6,7)	0.405	1	0.0030	0.388
(5,7)	(1,6)	0.532	11	0.0043	0.378
(6,7)	(2,7)	0.681	51	0.0027	0.377
(1,7)	(2,7)	0.495	59	0.0017	0.357
(1,6)	(4,4)	0.414	6	0.0039	0.353
(2,4)	(1,6)	0.559	19	0.0028	0.351
(7,3)	(7,1)	0.399	10	0.0036	0.342
(7,3)	(1,6)	0.660	33	0.0036	0.334
(7,6)	(1,6)	0.563	11	0.0065	0.314
(7,7)	(1,6)	0.691	11	0.0042	0.311
(3,1)	(2,7)	0.556	58	0.0022	0.307
(6,1)	(2,7)	0.566	54	0.0031	0.302
(7,1)	(2,7)	0.742	73	0.0022	0.301
(3,2)	(3,6)	0.402	1	0.0023	0.082
(6,3)	(6,7)	0.482	1	0.0052	0.299
(6,6)	(6,7)	0.477	1	0.0059	0.292
(4,3)	(6,7)	0.511	6	0.0022	0.286
(7,2)	(2,7)	0.766	73	0.0030	0.280
(1,2)	(6,6)	0.422	19	0.0019	0.280
(7,6)	(6,7)	0.450	1	0.0058	0.279
(5,7)	(2,7)	0.515	51	0.0032	0.278
(4,3)	(1,6)	0.584	19	0.0022	0.278

TABLE I

RAINFALL AND SURFACE WATER STAGE PATTERNS WITH ASSOCIATION RULE SUPPORT AND CONFIDENCE

Grid cell (row,column)		Root Mean Squared Error (RMSE)	
Rainfall	Stage	Linear Regression	Support Vector Regression
(6,1)	(1,6)	2169.68	3.89
(2,7)	(6,7)	1.57	1.40
(5,7)	(1,6)	679.02	3.55
(6,7)	(2,7)	75.26	19.64
(1,7)	(2,7)	75.88	3.30
(1,6)	(4,4)	4.02	4.05
(2,4)	(1,6)	3456.16	11.47
(7,3)	(7,1)	2.95	2.81
(7,3)	(1,6)	960.16	3.43
(7,6)	(1,6)	1611.31	3.61
(7,7)	(1,6)	936.22	5.08
(3,1)	(2,7)	78.42	1.61
(6,1)	(2,7)	87.02	4.52
(3,2)	(3,6)	3.123	2.77
(7,1)	(2,7)	76.68	6.25
(6,3)	(6,7)	1.64	1.37
(6,6)	(6,7)	1.85	1.46
(4,3)	(6,7)	1.70	1.48
(7,2)	(2,7)	76.04	4.29
(1,2)	(6,6)	3.02	3.00
(7,6)	(6,7)	1.97	1.55
(5,7)	(2,7)	78.84	1.75
(4,3)	(1,6)	1564.14	3.49

TABLE II

THE ROOT MEAN SQUARED ERROR VALUES FOR BOTH LINEAR AND SUPPORT VECTOR REGRESSION MODELS FOR EACH SPATIO-TEMPORAL PATTERN

Source	Degrees of Freedom	Sum of Squares	Mean Squares	F-statistic	Pr($\leq$F-statistic)
Rainfall	1	100.70	100.699	17.909	6.73e-05
Residuals	72	404.85	5.623		

TABLE III

ANALYSIS OF VARIANCE (ANOVA) TABLE FOR LINEAR REGRESSION OF THE PATTERN $RAIN_{(3,2)} \rightarrow WATERGAGE_{(3,6)}$.

Source	Degrees of Freedom	Sum of Squares	Mean Squares	F-statistic	Pr($\leq$F-statistic)
Rainfall	1	102.80	102.798	29.678	6.756e-07
Residuals	72	249.39	3.464		

TABLE IV

ANALYSIS OF VARIANCE (ANOVA) TABLE FOR LINEAR REGRESSION OF THE PATTERN $RAIN_{(6,6)} \rightarrow WATERGAGE_{(6,7)}$.

CHAPTER 7

CONCLUSION

In this thesis, a relationship is established between the rainfall and the surface water datasets. The datasets are mined and frequently occurring spatio-temporal patterns are discovered. This includes estimating the time delay between the occurrence of rainfall and the corresponding change in the surface water stage for each spatio-temporal patterns. The correlation, time lag, support and confidence of these patterns is show in Table I. These patterns are used to build regression models that are in turn used to predict the surface water stage based on the spatio-temporal pattern, the time delay and historical data. These predictions are verified by evaluating them with the corresponding actual surface water stage values using the Root Mean Squared Error (RMSE) metric as shown in Table II. Hence, there is a relationship established between rainfall and surface water and rainfall data can be used to estimate the surface water stage values.

CITED LITERATURE

1. Huang, Y., Shekhar, S., and Xiong, H.: Discovering colocation patterns from spatial data sets: a general approach. Knowledge and Data Engineering, IEEE Transactions on, 16(12):1472–1485, 2004.

2. Celik, M., Shekhar, S., Rogers, J. P., Shine, J. A., and Yoo, J. S.: Mixed-drove spatio-temporal co-occurence pattern mining: A summary of results. In ICDM, volume 6, pages 119–128, 2006.

3. Murphy, K. P.: Machine learning: a probabilistic perspective. MIT Press, 2012.

4. Harms, S. K., Deogun, J., and Tadesse, T.: Discovering sequential association rules with constraints and time lags in multiple sequences. In Foundations of Intelligent Systems, pages 432–441. Springer, 2002.

5. Patel, D.: Interval-orientation patterns in spatio-temporal databases. In Database and Expert Systems Applications, pages 416–431. Springer, 2010.

6. Changnon, S. A., Huff, F. A., and Hsu, C.-F.: Relations between precipitation and shallow groundwater in illinois. Journal of Climate, 1(12):1239–1250, 1988.

7. Bartlett, M. S.: The statistical analysis of spatial pattern. Springer, 1976.

8. Cruz, I. F., Ganesh, V. R., Caletti, C., and Reddy, P.: GIVA: A Semantic Framework for Geospatial and Temporal Data Integration, Visualization, and Analytics. In Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, pages 534–537. ACM, 2013.

9. Caletti, C.: Ontology Matching Enhanced with Similarity Measures for Georeferenced Datasets. Master's thesis, University of Illinois at Chicago, 2014.

10. Benesty, J., Chen, J., and Huang, Y.: Time-delay estimation via linear interpolation and cross correlation. Speech and Audio Processing, IEEE Transactions on, 12(5):509–519, 2004.

11. Agrawal, R., Srikant, R., et al.: Fast algorithms for mining association rules. In Proc. 20th int. conf. very large data bases, VLDB, volume 1215, pages 487–499, 1994.

12. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2013. ISBN 3-900051-07-0.

13. Paradis, E., Claude, J., and Strimmer, K.: APE: analyses of phylogenetics and evolution in R language. Bioinformatics, 20:289–290, 2004.

14. Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D., and Weingessel, A.: Misc functions of the department of statistics (e1071), tu wien. R package, pages 1–5, 2008.

15. Liu, B.: Web data mining. Springer-Verlag Berlin Heidelberg, 2007.

16. Cheng, M., Yu, P., and Liu, B.: Advances in Knowledge Discovery and Data Mining: 6th Pacific-Asia Conference, PAKDD 2002, Taipei, Taiwan, May 6-8, 2002. Proceedings. Number v. 6 in Lecture Notes in Artificial Intelligence. Springer, 2002.

17. Sclove, S.: A Course on Statistics for Finance. Taylor & Francis, 2012.

VITA

Pavan Reddy

Degrees	B.E., Visvesvaraya Technological University, 2007 M.S., University of Illinois at Chicago, 2014
Publications	**GIVA: A Semantic Framework for Geospatial and Temporal Data Integration, Visualization, and Analytics** I. F. Cruz, V. R. Ganesh, C. Caletti, and P. Reddy ACM SIGSPATIAL, November 2013