PhenomenaAssociater: Linking Multi-Domain Spatio - Temporal datasets;

Prathamesh Walkikar[1], Vandana P. Janeja[1]

[1] University of Maryland Baltimore County, Baltimore, USA

Abstract. This paper focuses on the demonstration of an analytics dashboard application for analyzing interesting spatio-temporal associations between anomalies across multiple spatio-temporal datasets, potentially from disparate domains, to find interesting hidden relationships. The proposed system is intended to analyze spatio-temporal data across multiple phenomena from disparate domains (for example traffic and weather) to identify interesting phenomena relationships by linking anomalies from each of these domain datasets. This web-based dashboard application developed in R Shiny provides interactive visualizations to quantify the multi-domain associations. The application uses a novel framework of algorithms and quantification metrics to associate these anomalies across multiple domains using spatial and temporal proximity and influence metrics.

1 Introduction

In today's world, it is not a surprise to find that almost everything in this world is inter-related particularly nearby things. These relationships also become fundamentally true across multiple application domains. For example, (a) Weather condition at a location will impact traffic [2], (b) Oil spills in oceans will adversely impact underlying aquatic animal population [3], (c) pollution in a location can affect disease spread and many more [4]. The common link across phenomenon is the underlying geographical space, which can help associate phenomena in a particular region to link with other inter-related phenomenon in the same region. This can be in the form of spatio-temporal associations. This preposition however faces numerous obstacles in the form of tremendous amount of single domain data and difficulty in combining data across different domains due to data heterogeneity issues.

These data integration and heterogeneity issues can be avoided by looking at extracted knowledge from individual phenomena and then look for potential associations across the discovered knowledge capturing data for a phenomenon, instead of looking at raw data form distinct domains. The extracted knowledge in each domain in our case is the anomalous window comprising of a set of contiguous points in a region that are unusual with respect to the rest of the points in the region. One example of multi-domain anomaly detection is discussed in [5] where circle based scan windows from single domains are linked using traditional spatial associations. In this paper, we propose a novel dashboard to illustrate the discovery of such multi-domain anomalies through novel influence metrics which quantify the associations between phenomena in both spatial and spatio-temporal datasets. Our web-based dashboard application discovers these single domain anomalies across individual domain datasets and associates them to derive interesting relationships between different domains capturing multiple phenomena associations.

2 Demonstration

Developed in R, this dashboard application can assist domain experts in deriving potential knowledge from multi-domain spatio-temporal datasets and thus, facilitate researchers studying impact analysis of one phenomena over others in research fields like epidemiology, traffic accident analysis, impact of wildfires [6] to name a few. Figure 1 shows a sample of association results in the form of detected anomalous windows overlayed on a map visualization for a real world spatio-temporal health-ranking outcome dataset where we discover interesting spatio-temporal associations using influence indicators between child poverty rates and unemployment rates in the State of Maryland, USA.

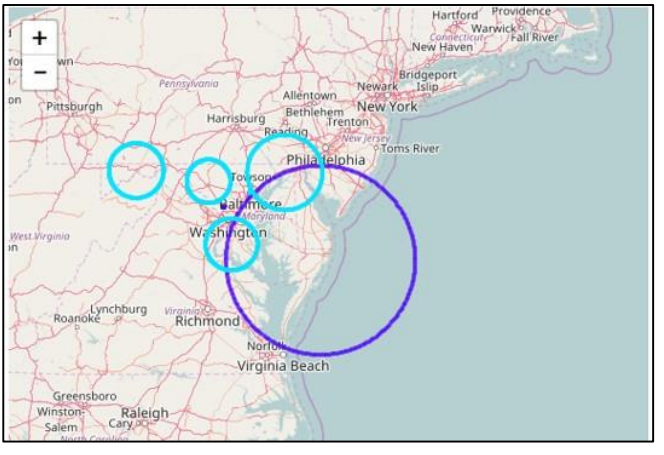

Figure 1: Multi-domain anomalous association results

In the demo, we will show step by step how PhenomenaAssociater discovers interesting associations between such multi-domain spatio-temporal datasets.

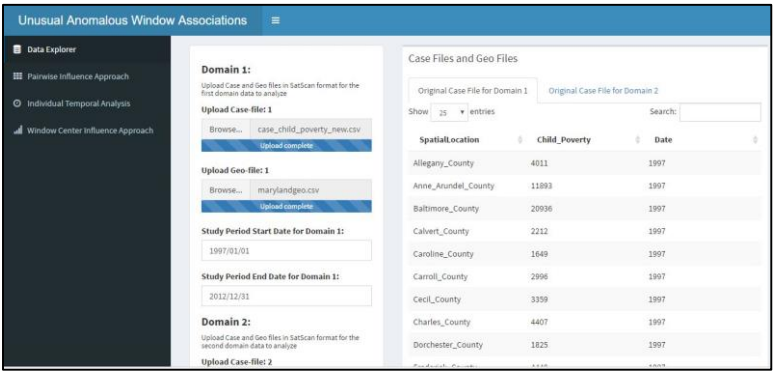

Figure 2. Data pre-processing and management component

2.1 Discovering single domain spatio-temporal anomalies

Spatio-temporal data handling and management forms an important component in our analysis process. For handling the temporal elements of our data, we discretize the data by time and then perform anomaly detection in each of the temporal bins created. Our application is equipped with three different categories of data discretization techniques which include - equal frequency binning, equal width binning and hierarchical time series clustering based binning strategies. Data Modelling component handles the generation of specific input files from individual discretized instances of datasets, which are inputs for associations, such as specific formats required for SatScan [7]. Figure 2 depicts this spatio-temporal data modelling and management view of our application.

By using existing anomalous window discovery techniques such as space-time scan statistics using SatScan, SSLIP [8] and RWSCAN [9], single domain anomalous windows can be extracted from spatio-temporal datasets for each domain such as child poverty or unemployment data. Our interface is much more tightly coupled with SatScan mainly due to its wide use and intuitive findings.

2.2 Anomalous Window Associations

After these single-domain anomalous windows have been extracted, we associate these windows based on proximity and overlap patterns of these anomalous windows discovered from the single domain anomaly detection methods. In other words, co-occurrence of anomalous windows from different domains in the same geographical areas of proximity over time determines spatio-temporal association. We propose the concept of influence distance, which quantifies the overlap across the phenomena both in terms of spatial and non-spatial attributes. We also propose and utilize influence score - a novel metric for measurement of spatio-temporal associations as well as variations of spatio-temporal confidence and support and lift measures which quantify these interesting associations. Due to limited scope of this paper we mainly define influence distance:

DEFINITION 1. *[Influence distance] Let v be the given phenomenon, and p and q be two spatial objects. We define* $d^{v}_{p\to q}$ *as the* ***influence distance*** *from spatial object p to q for the phenomenon v.* $d^{v}_{p\to q}$ *is the sum of the weights of the constituent edges of the shortest path from p to q in the network of v.*

Hence, if p and q are one spatial object, then $\mathrm{d}^{\mathrm{v}}_{\mathrm{p}\to\mathrm{q}} = 0$. If p and q are not connected, then we get the influence distance $\mathrm{d}^{\mathrm{v}}_{\mathrm{p}\to\mathrm{q}} = \infty$.

Here network of phenomena is derived using a spatial neighborhood [15] approach. Based on influence distance, we calculate the influence score, which quantifies the proximity and overlap of anomalous windows, where we take the aggregate influence distance between the spatial nodes present within each of the distinct domain windows. When a user wants to analyze a set of two distinct inter-related domains, after running anomalous window discovery methods, the user gets a series of single domain anomalous windows with respect to time for each distinct domain. To further associate these anomalies, our application provides the user to choose from a set of distinct approaches

to effectively associate these domain anomalies to find interesting phenomena associations. These links are quantified using the distinct set of influence indicators which are plotted in Figure 3.

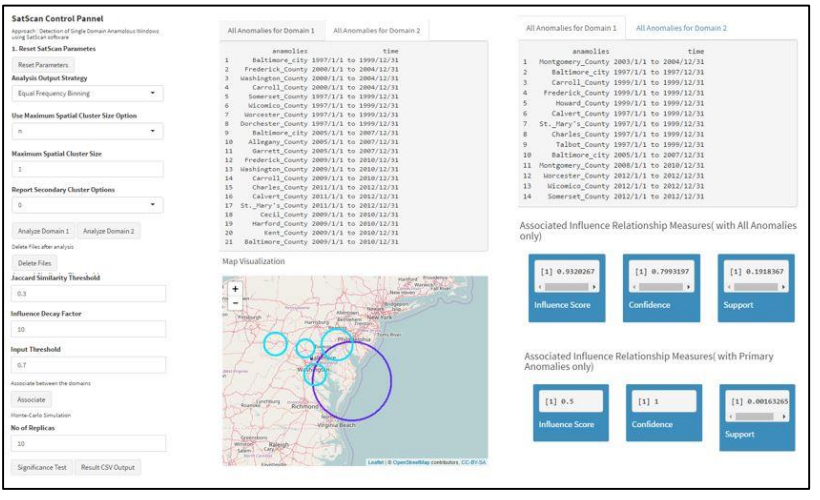

Figure 3. Influence relationship results

These results can act as a supplementary information for domain experts in order to obtain useful phenomena linkages to further study rare phenomena relationships.

Acknowledgement

This work is supported in part by the US Army Corps of Engineers, agreement number: W9132V -15-C-0004.

References

1. Winston Chang, Joe Cheng, J.J. Allaire, and Jonathan McPherson. 2015. shiny: Web Application Framework for R. (2015).
2. L.S. Nookala. 2006. *Weather Impact on Traffic Conditions and Travel Time Prediction*. thesis.
3. M. Barron. 2011. Ecological Impacts of the Deepwater Horizon Oil Spill: Implications for Immunotoxicity. (2011).
4. "Climate Sensitive Diseases"
5. R.P. Costa Mário, Liliana Caramelo, Carmen Vega Orozco, and Mikhail Kanevski. 2013. Assessing SatScan ability to detect space-time clusters in wildfires. *EGU General Assembly 2013* (2013).
6. V. P. Janeja, N. Adam, V. Atluri, J.S. Vaidya. Spatial neighborhood based anomaly detection in sensor datasets, Data Mining and Knowledge Discovery, special issue on Outlier Detection, 20(2), March 2010, pp. 221-258, Springer.
7. Kulldorff M. and Information Management Services, Inc. SaTScanTM v8.0: Software for the spatial and space-time scan statistics. http://www.satscan.org/, 2009
8. L. Shi and V. P. Janeja. 2009. Anomalous window discovery through scan statistics for linear intersecting paths (SSLIP). Proc. of the 15th intl. conf. on Knowledge discovery and data mining - KDD '09.
9. V.P. Janeja and V. Atluri. 2008. Random Walks to Identify Anomalous Free-Form Spatial Scan Windows. IEEE Transactions on Knowledge and Data Engineering 20, 10 (2008), 1378–1392.
10. V.P. Janeja and R. Palanisamy. 2012. Multi Domain Anomaly Detection in Spatial Datasets, Knowledge and Information Systems Journal, DOI: 10.1007/s10115-012-0534-5, (2012).