Time Series Projection to Highlight Trends and Outliers

Eren Cakmak Daniel Seebacher Juri Buchmüller Daniel A. Keim

Data Analysis and Visualization Group
University of Konstanz*

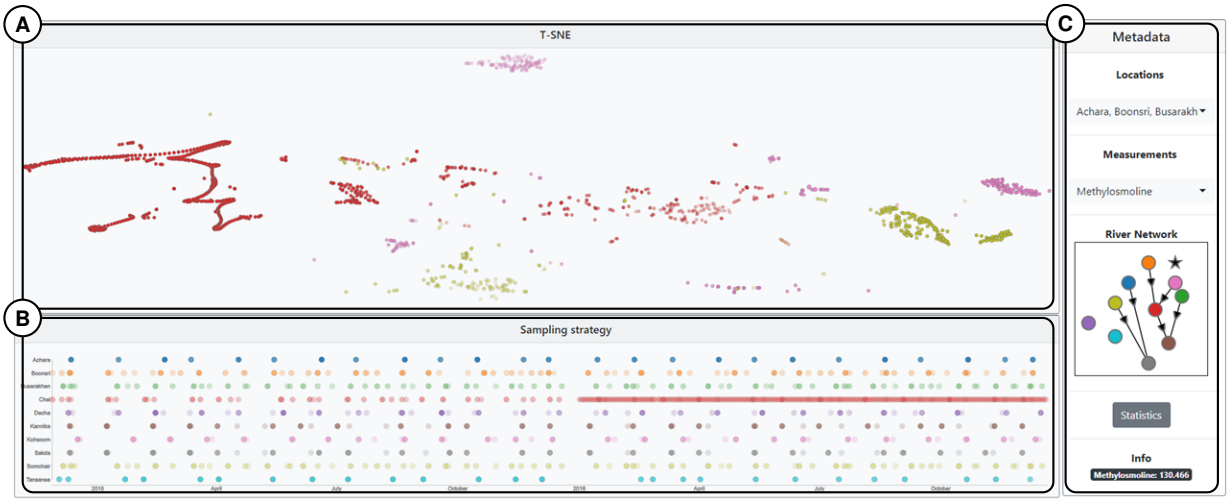

Figure 1: ViCCEx – Visual Chemical Contamination Explorer (A) The t-SNE projection enables to identify patterns, trends, and outliers in multivariate time series sensor readings (B) Sampling strategy view to investigating the chemicals measurements taken at each location (C) Metadata panel with filter options and further extracted statistics for each chemical for each sensor location.

Abstract

The goal of the VAST Challenge 2018 Mini Challenge 2 (MC 2) was to unveil the possible causes and effects of environmental pollution in the Boonsong Lekagul Wildlife Preserve. We propose the ViCCEx (Visual Chemical Contamination Explorer) system that enables to interactively explore the sparse multivariate river network sensor reading dataset to identify characteristics, trends, and outliers of the different sensor reading locations over time. The ViCCEx system uses a t-SNE projection to display an overview visualization, a sampling strategy view to highlight the overall sampling strategies of different chemical measurements at each sensor location, and various extracted statistics to highlight the evolution of chemical measurements. The three views are connected via linking and brushing, which enables to explore and identify possible pollution causes and effects in the preserve.

Index Terms: Human-centered computing—Visualization—Visualization application domains—Visual Analytics;

1 Introduction

MC 2 of the VAST Challenge 2018 outlines an uncertain situation on environmental impacts in the nature preserve known from 2017. To determine the extent of the pollution, experts turn to the analysis of the water quality in the preserve's rivers. Almost two decades of infrequent samples in different river networks and measuring stations for more than 100 individual chemicals have to be analyzed. Besides a thorough investigation into trends and anomalies in the measured chemicals, the focus lies as well on analyzing the sampling strategies and the impact on the infamous Rose-Crested Blue Pipit. The combination of a large volume of irregularly sampled, sparse time series with more than 100 features over a large period of time with spatial references demands for a Visual Analytics [2] tool guiding a user in the exploration of the data.

We introduce ViCCEx, the Visual Chemical Contamination Explorer[1]. ViCCEx uses a projection approach for time series further described in section 3 to allow the identification of unusual behavior in time series, as well as additional views, to investigate the sampling strategy and additional statistics about the chemical measurements. The views on the projection, temporal sampling strategy, and the development of specific values are linked together. The ViCCEx principle can be applied to time series in general to reveal trends and outliers. In the following sections, we present our approach and demonstrate how ViCCEx can be applied to the provided challenging dataset following a Visual Analytics approach to identify potential suspicious patterns.

2 Data

The data provided are measured values from ten measuring stations in the Boonsong Lekagul Wildlife Preserve. The period of the measurements extends from January 1998 to December 2016. In total, more than 136,000 measurements of 106 different chemicals were carried out at the ten stations. Providing a multivariate time series dataset for each of the ten stations. However, one problem is that not every chemical is measured at every station at every point in time, which is why the dataset is very sparse. Further, the individual measuring sensor locations are connected in a river network which means that the contamination of one sensor station can influence the chemical contamination of another sensor location.

*e-mail: firstname.lastname@uni-konstanz.de

[1] https://viccex.dbvis.de/

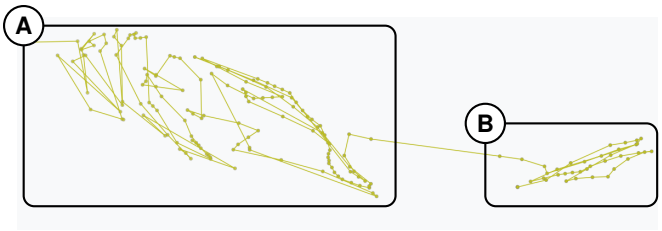

Figure 2: Somchair sensor t-SNE projection (A) measurements taken from 11/2009 until 11/2015, and in (B) from 1/2016 until 12/2016. The two clusters indicate that between (A) and (B) there was a change in the chemical contamination in this area. We assume that there was environmental pollution in this area in December 2015 since there was an increase in the chemical Methylosmoline.

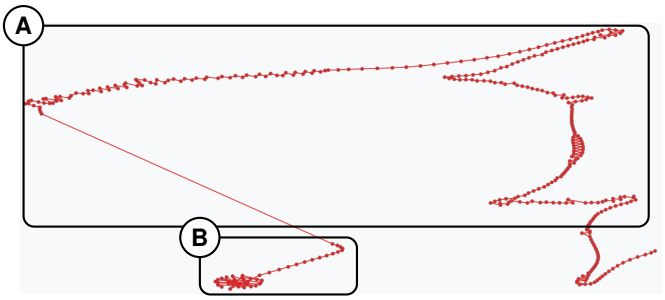

Figure 3: Chai station t-SNE projection from 01/2016 until 12/2016 (A) Small constant differences between the measurements caused mostly by the daily water temperature measurements (B) In November 2016 multiple chemicals decreased.

3 Analysis Approach

For the analysis of MC 2, we decided to follow the Visual Analytics mantra of Keim et al. [2]: "Analyse First – Show the Important – Zoom, Filter and Analyse Further – Details on Demand". In following this mantra, we show the analyst in a first overview visualization (see Figure 1 (A)) what is important: strong fluctuations in the measured chemicals, as these, are an indication of a drastic change in the chemical contamination of the wildlife preserve. In a first step, we prepare the data by transforming it into a 106-dimensional feature vector with the chemical measurements for each station and each day. The dataset is relatively sparse, since not every chemical is measured every day, which is why we use linear interpolations between a previously and a subsequently measured value to fill in missing values.

We apply the t-SNE projection technique to this dataset to generate an overview. Further, we color each sensor location in the t-SNE using a unique color. Here we follow a similar approach to Bach et al.'s Time Curves [1]. If two measurements are relatively similar, they are drawn close together in the projection, if they are dissimilar, however, far apart. The difference between the two measurements should theoretically be relatively small, for example, slight changes in water temperature, which results in them being drawn closer together in the t-SNE projection. However, chemical contamination should cause some of the readings to change significantly, resulting in a significant distance between two consecutive readings, which is why they are drawn far apart in the projection. Such clusters can be seen in our overview visualization in Figure 1. For some stations, a relatively dense cluster of measured values indicates no unexpected behavior of the readings. For other stations, however, two or three clusters can be seen, which shows that the measured values have changed several times drastically.

The t-SNE overview visualization serves as a starting point for further analyses. The projection can be used, for example, to select stations with drastic changes. These can then be further analyzed using the sampling strategy view as well as extracted statistics and the time series view. Two examples illustrate how the t-SNE view can help in the analysis of chemical measurements. First, Figure 2 shows our t-SNE time curve for the Somchair station. Here it can be clearly seen that between November 2015 (A) and January 2016 (B) there is a huge difference in the measured chemical concentrations. The unusual difference in the concentration may be a clear sign of an external influence, e.g. an illegal waste dumping.

Second, Figure 3 shows an interesting pattern for the station Chai. First, the line is relatively long, with small permanent changes in the t-SNE projection. Measurements were carried out daily for Chai after January 2016, which causes the line effect. Often only water temperature was measured for each day which can be examined in more detail in the sampling strategy view. Nevertheless, a break can be seen again, which also occurs, as for Somchair, around November 2016 (Figure 3 (A-B)). This find raises the suspicion that the reason for the changes is artificial due to the magnitude of the changes. A possible explanation could be that chemicals were removed from the Chai area.

To identify and understand patterns as well as anomalies of the sampling strategy at each location, we provide a sampling strategy view. The timeline visualization shows each site in a unique coloring and is linked to all other views. The time interval can be adapted in this visualization using zooming and panning. Further, locations and chemical measurements can be filtered using either an on click on an individual data point or by using the metadata panel. An aggregation function enables to reduce to the number of depicted data points in the sampling strategy view.

The first two visualizations (see Figure 1(A, B)) enable an overview on the dataset to identify and filter interesting time intervals and samples taken at the different sensor locations. The statistics and line chart view aid to find patterns, trends, and outliers in individual chemical measurements at each location. For this, we compute 17 statistics (e.g., median, variance, standard deviation, etc.) and sort the already filtered chemicals in a table view. The table view highlights individual chemicals with unusual variations and trends. For instance, the mean change can indicate the overall trends of increasing or decreasing values. After selecting one chemical in the statistics view, a line chart is displayed that shows the evolution of the chemical measurements at each location. The line chart allows analyzing the chemical trends, patterns, and outliers at each location. Furthermore, the line chart view helps to detect correlations between different sensor stations.

4 Conclusion

The ViCCex system enables to explore and analyze sparse multivariate time series data interactively. We were able to solve the task of the challenge and detect two contaminated areas in the wildlife preserve using the proposed system. Furthermore, we identified multiple trends, patterns, and outliers. Using the system, we were able to solve the tasks of the challenge. Our main contribution is the interactive exploration of the multivariate time series data using a t-SNE projection, sampling strategy, a line chart, and various extracted statistics.

References

[1] B. Bach, C. Shi, N. Heulot, T. Madhyastha, T. Grabowski, and P. Dragicevic. Time curves: Folding time to visualize patterns of temporal evolution in data. *IEEE transactions on visualization and computer graphics*, 22, 2016.

[2] D. A. Keim, F. Mansmann, J. Schneidewind, J. Thomas, and H. Ziegler. Visual analytics: Scope and challenges. In *Visual data mining*, pp. 76–90. Springer, 2008.