Identifying Patterns and Anomalies within Spatiotemporal Water Sampling Data

VAST Challenge 2018: Award for Elegant Design of an Interactive Display

Isabel Piljek Giuliana Dehn Jannik Frauendorf Ziad Salem Yerzhan Niyazbayev
Juri Buchmüller Eren Cakmak Wolfgang Jentner Florian Stoffel Daniel A. Keim

University of Konstanz, Germany*

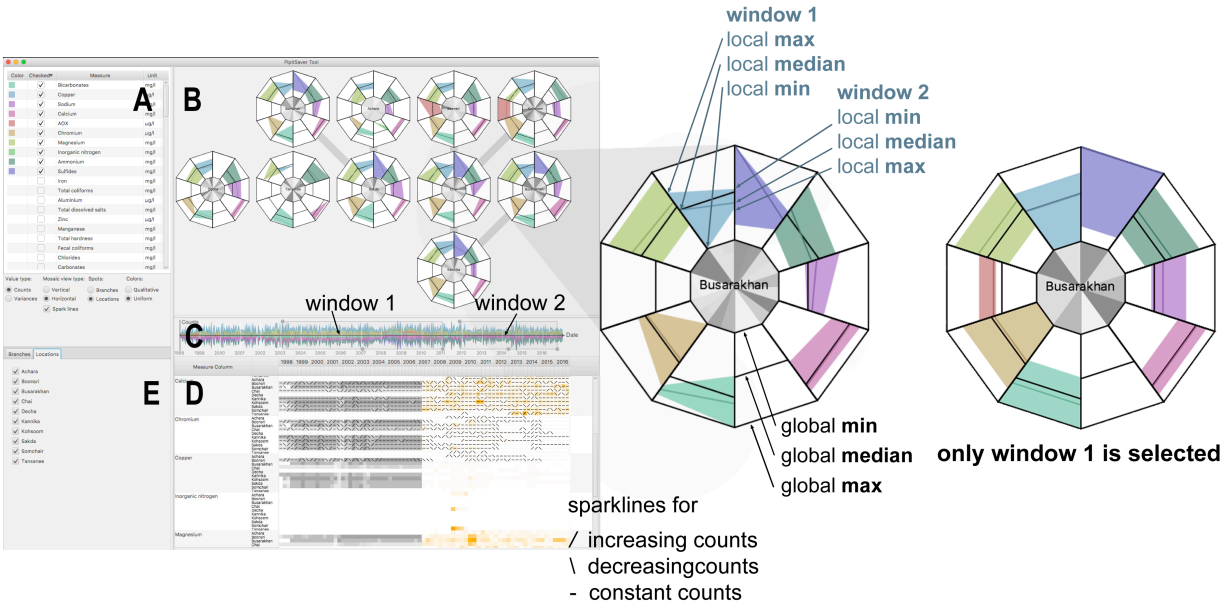

Figure 1: Water sampling data of the VAST Challenge 2018 displayed with our custom system. The selected measures (A) are linked to the *crystal objects* (B), the *stream time graph* (C), and the *mosaic plots* (D). The user can further streamline the analysis by (de-)selecting specific locations or branches (E).

Abstract

This paper presents our solution to the Mini Challenge 2 (MC2) of the VAST Challenge 2018. We will analyze the provided data set and introduce our visualization tool, which was implemented and tailored to the tasks given by MC2. The tool combines the power of stream graphs, innovative glyph visualizations, box plots, sparklines, heat maps and cross-filter strategies. It allows identifying patterns and anomalies within the provided data set.

1 Introduction

The VAST Challenge 2018 is about finding evidence for the population decline of the blue pipit bird species in a nature preserve. MC2 provides water sampling data from different locations within this preserve. The task consists in characterizing the data and finding trends, as well as anomalies, with respect to contamination in the preserve's waterways. Furthermore, the sampling strategy is to be critically examined.

*e-mail for all authors: {*firstname.lastname*}@uni-konstanz.de

2 Data Preprocessing

The *Mistford Hydrology Department* collected water samples from ten different locations, which can be grouped into four river branches. The samples cover 106 measures including the sample date, value, unit, and location. The data is preprocessed using KNIME [1] and SQL and is transformed and stored in a data warehouse in a PostgreSQL[1] database management system. The entire time span covers 19 years, but only 28.3% of this period contains records. The data is sampled at irregular time intervals and shows big gaps. Using the data warehouse and analyzing the data set with Tableau[2], we identified duplicated entries having the same measure on the same date, at the same location, with the same values. These duplicates make up 24.04% of the total data set.

3 Application

In this Section, we present our prototype that supports an in-depth analysis of the data through the use of Visual Analytics [3]. Figure 1 shows the interface of the system. The selected measures (A) are linked to the *crystal objects* (B), the *stream time graph* (C), and the *mosaic plots* (D). The user can further streamline the analysis by (de-)selecting specific locations or branches (E).

[1] https://www.postgresql.org/
[2] https://www.tableau.com/

Crystal View The crystal view depicts box-plot similar visualizations for each selected measure in a different color. The values are scaled on a logarithmic base and refer to the two time windows of the stream time graph. An empty section indicates that in at least one of the time windows no samples are available. If one of the two time windows is deactivated the *crystal view* updates accordingly as shown in Figure 1 (only window 1 is selected). Using both time windows, the user can quickly observe trends over time for multiple measures and compare the behavior of these measures in different locations. The normalized variances are mapped onto shades of gray in the glyph's center. A darker gray indicates a higher variance and thus, a possibly more interesting feature. The glyphs are connected with gray lines representing the water flow and the river branches.

Stream Time Graph The stream time graph is a variation of a Stacked Area Graph depicting time on the x-axis and sampling counts on the y-axis. In contrast to the usual Stacked Area Graphs, this so-called Stream Graph [2] displaces values around a varying central baseline. As a result, the distance of each area to the base line is minimized, which improves the trackability of the areas.

Mosaic Plot An overview of the sample counts of the measures can be seen in the mosaic plot. This view reveals groups of measures having a similar sampling pattern. Similar patterns indicate similar sampling strategies, as shown in Figure 2. The mosaic plot presents either sample counts (gray) or variances (orange). These are sequentially mapped to the color. The values of the mosaic plot are aggregated into quarters to provide an overview of the entire time span. By default, a row represents one location. However, the values of one river branch can be aggregated to reduce visual clutter. It is also possible to deselect locations or river branches to tailor the analysis. The sparklines overlay the background color and refer to the number of records being obtained. We discretize these continuous values by plotting whether the number of records increased, decreased, or stayed equal with respect to the previous quarter. This makes an investigation of patterns and anomalies in the sampling strategy possible.

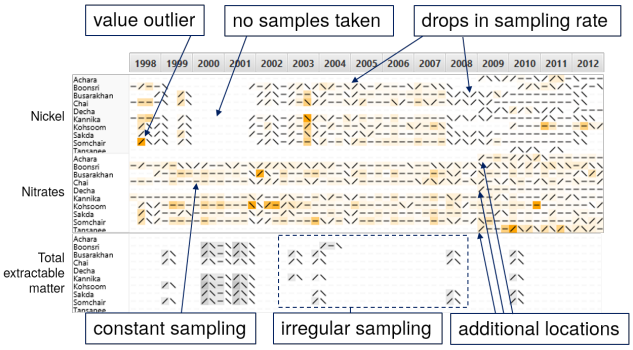

Figure 2: Visible patterns in the mosaic plot: Values of measures showing either high variances (orange) in different time spans or counts (gray). Sparklines indicate the trend of the number of records with respect to the previous quarter.

4 Analysis

Our customized system allows the user to apply various filters and options to support the analysis task with the help of Visual Analytics. The next paragraphs show this by using the example of the pollution analysis task.

By looking at the mosaic plot (Figure 1D or 2) the user gets a first impression of the sampled measures over the entire period of time. The plot reveals when, and how regular samples were taken. In 2009, new measuring stations where added in *Achara*, *Decha* and *Tansanee* and the coverage of many measures improved. Plotting variances illustrates that methlyosmoline and AGOC-3A have high variances starting in 2016, especially in *Kohsoom*. The mosaic plot furthermore shows groups of measures with similar sampling behavior at a glace (Figure 1D calcium and chromium). The user further narrows down the search for locations or river networks (Figure 1E) using the provided filters. After selecting individual measures (Figure 1A), the stream time graph (Figure 1C and 3) shows the exact number of records per measure and functions as a time filter for the crystal objects (Figure 1B and 3). After adjusting the windows to the interval of interest, we see a drastic increase of methylosmoline in *Kohsoom* and *Somchair*, whereas it decreases in the remaining locations. The highest values of methylosmoline over the entire period of time were measured in *Kohsoom* and *Somchair*. A combination of the system's capabilities and the input of the user make a wide range of analysis possible.

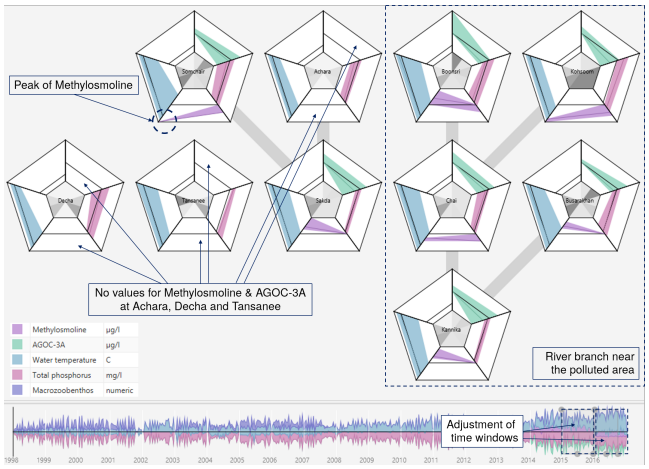

Figure 3: The crystal glyphs allow the analysis of spatial distribution of selected measures. The marked river branch on the right around Kohsoom is clearly contaminated with methylosmoline. Also, the absence of samples in different locations (left) is visible, and trends can be compared.

5 Conclusion

We present an innovative system helping to answer the questions of the VAST Challenge 2018 MC2 through Visual Analytics. Our tool can be used to detect patterns and anomalies in the sampling strategy as well as supporting the analysis of sample values. We can further investigate trends and compare these trends across multiple features and locations. Cross-filters and aggregations help tailor the analysis and reduce visual clutter. The various visualizations and interaction possibilities enable the user to perform a diverse set of analysis tasks.

References

[1] M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, P. Ohl, C. Sieb, K. Thiel, and B. Wiswedel. KNIME: The Konstanz Information Miner. In *Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)*. Springer, 2007.

[2] L. Byron and M. Wattenberg. Stacked graphs–geometry & aesthetics. *IEEE transactions on visualization and computer graphics*, 14(6), 2008.

[3] D. Keim, G. Andrienko, J.-D. Fekete, C. Görg, J. Kohlhammer, and G. Melançon. Visual analytics: Definition, process, and challenges. In *Information visualization*, pp. 154–175. Springer, 2008.