# Matrix-Based Visual Correlation Analysis on Large Timeseries Data

Michael Behrisch* Universität Konstanz
James Davey† Fraunhofer IGD
Tobias Schreck‡ Universität Konstanz
Daniel Keim§ Universität Konstanz
Jörn Kohlhammer¶ Fraunhofer IGD

**ABSTRACT**

In recent years, the quantity of time series data generated in a wide variety of domains grown consistently. Thus, it is difficult for analysts to process and understand this overwhelming amount of data. In the specific case of time series data another problem arises: time series can be highly interrelated. This problem becomes even more challenging when a set of parameters influences the progression of a time series. However, while most visual analysis techniques support the analysis of short time periods, e.g. one day or one week, they fail to visualize large-scale time series, ranging over one year or more. In our approach we present a time series matrix visualization that tackles this problem. Its primary advantages are that it scales to a large number of time series with different start and end points and allows for the visual comparison / correlation analysis of a set of influencing factors. To evaluate our approach, we applied our technique to a real-world data set, showing the impact of local weather conditions on the efficiency of photovoltaic power plants.

**Index Terms:** H.3.3 [Information Search and Retrieval Design Tools and Techniques]: Information filtering—

## 1 INTRODUCTION

Large amounts of time series data are produced in a wide variety of domains, and their visual analysis is an important problem. A common analysis task is the comparison of measurements over time, with the aim of discovering meaningful correlations between measurements. An example is the correlation of temperature and power consumption measurements in an energy supply / consumption context. Many existing visual analytics techniques are restricted to the comparison of short time periods (e.g. weeks or days) and do not scale well for longer time series. We present a technique for the visual correlation analysis of numerous, potentially interrelated time series datasets. Our technique centers on a matrix representation and allows for the simultaneous comparison of multiple, overlapping time series of varying length. We illustrate the principle of our technique by applying it to a real-world data set, including measurements of weather conditions and performance parameters of nodes in a power grid. We also discuss interesting options for future work.

## 2 RELATED WORK

The analysis of time series data is a common problem in many application domains. Basic analysis goals include finding the minimum, maximum and average values in a specific time series, and the identification outliers. More complex analysis goals include clustering large sets of time series, extracting frequent time series patterns, and similarity searches with the help of user specified time series. Fundamental analysis techniques for representing, indexing and correlating time series data are discussed in [4]. Time series clustering techniques are surveyed in [6]. Visualization of a set of time series data is another field of research, surveyed in [1]. It requires a scalable representation that (1) supports a large number of potentially correlating time series datasets and (2) takes the length of time-and-sampling-resolution-dependent vectors into account.

Our work is also related to interactive matrix visualization, as a promising approach for scalable representation of large-scale network relationships. A large-scale, interactive graph visualization is presented by Elmqvist et al. in [2]. It provides a zoomable interface for an overview first / detail on demand [5] approach. In addition, every cell of the matrix contains an adaptive glyph representation, showing statistical features of the underlying network. Ziegler presented in [7] a financial performance chart growth matrix with the goal of identifying the best entry and holding times of unrelated / independent financial bonds.

Our time series matrix visualization differs from related work in two main points. Firstly, we visualize a large number of time series with different start and end points in one matrix. Secondly, it allows for a visual comparison / correlation analysis of different time series datasets in a comprehensive zoomable interface that enables the exploration of underlying correlation effects.

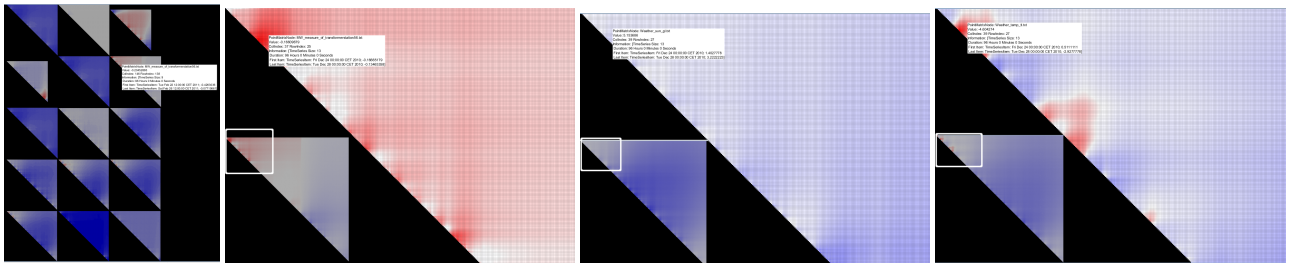
## 3 POWER GRID DATASET AND ANALYSIS PROBLEM

In collaboration with a German national energy provider, we considered two real world data sets obtained from the town of Freiamt, Germany [3]. Freiamt is home to a large number of regenerative power sources, including 160 small, roof-mounted photovoltaic, 3 biomass, and 3 small hydroelectric power plants. Up to 11 photovoltaic plants are connected to one *substation*, acting as a gateway to the power grid. In total, 29 substations and a weather observation station in Freiamt provide the measurement readings. The substation measurements are aggregations of the power generation and -consumption of multiple households and regenerative power plants. The weather station delivers a large number of weather parameters, e.g. rain fall rate, sunshine duration, temperature- and wind measurements, and visibility ranges (fog). In total, we consider nine weather parameters. The measurements span a duration from *2010-12-15* to *2011-12-17*, and are taken at intervals ranging between 10 and 30 minutes.

The goal of the analysis is to understand the interaction (correlation) between the state of the power grid and weather conditions. From a large number of potentially dependent measurement parameters, we want to find the parameter subset which is most useful for the analysis. While this is a problem in itself, it becomes even more challenging due to the fact that the correlations are inherently local with respect to scale and the time interval. These factors need to be considered in the visual analysis. Accordingly, one can find time-correlating predicate conditions, such as high temperature and long sunshine duration, that could lead to a drop or rise in the efficiency of the power grid.

The following data analysis questions arise: (1) Which parameters lead to a correlation between the reference and other time series datasets? (2) Can we show the parameter's impact on a reference dataset? (3) Which large-scale trends can be determined in long time series (e.g. in one year with more than $365 \times 24 \times 4$ sampling points)?

*e-mail: michael.behrisch@uni-konstanz.de
†e-mail: james.davey@igd.fraunhofer.de
‡e-mail: tobias.schreck@uni-konstanz.de
§e-mail: daniel.keim@uni-konstanz.de
¶e-mail: joern.kohlhammer@igd.fraunhofer.de



(a) Matrix Overview (b) Transformer Substation 56: MW/h (c) Sun Solar Radiation: Joule/cm$^2$ (d) Temperature in °C

Figure 1: (a) shows an overview of all time series matrices. (b) shows the efficiency of substation 56 in the selected time period. (c) and (d) reveal the impact of the sun's solar radiation and temperature on the substation's efficiency.

## 4 Timeseries Matrix Visualization

Our approach centers on a matrix visualization, as it can represent large numbers of time series in a pixel-oriented way, mapping each value to the color of a pixel. In this triangular matrix representation, the horizontal and vertical axes describe the start and endpoints of a specific time interval in the overall time series. Accordingly, each matrix point $x_{(i,j)}, i > j$ refers to a time series interval starting at time $t(i)$ and ending at time $t(j)$. To foster correlation analysis in the matrix we show statistics computed over the respective time series intervals, thus providing a tool for the *screening* of correlations at different intervals and offsets. The statistic values are presented in a mouseover tooltip.

The color of each data point represents a statistical measure $f_{(i,j)}$ computed over the interval $[t(i),t(j)]$. The measure, $f$, can be set by the user on-demand. Example measures include the trend (slope of the regression line), the standard deviation, average, or the geometric mean. Minimum, maximum, variance, sum and squared sum statistics can also be computed. The time series can be resampled on demand to set hours, days, months, etc. as the base unit of measurement.

The color map is an important design factor for the comparability of the matrices. Since, the transformer stations can have a positive or negative net output to the energy grid, depending on whether power is consumed or produced, we chose a bipolar red to blue color map. As Figure 1 (b) depicts, negative net outputs (power production states) are visually outstanding, due to their red color. Other measurement parameters are shown by additional triangular matrices in a small-multiples display. They represent the available weather information (e.g. air temperature 5 cm above ground level in °C, sunshine duration and rainfall rate) over the available time period. A local (per matrix) color map can be applied on demand to reveal the feature's special characteristics. A semantic zoom interface lets the user explore the correlations between the matrices in an overview or inspect the data characteristics in an emphasized time span. For the lowest semantic zoom level a line chart representation of the corresponding time span is planned.

## 5 Use Case

Figure 1 showcases one of the findings we made using our approach on the Freiamt dataset. Figure 1 (a) shows an overview over six transformer substations and nine weather parameters. Figure 1 (b) represents the positive and negative megawatt consumption rate of substation 56 in a larger view. This substation is especially interesting, since eleven photovoltaic power plants are connected to it. The visual task is to find power injection phases represented by a dark red color. Two visually outstanding areas exist in the one year time period. The first ranges from *2010-12-24* to *2010-12-26* and the second from *2011-02-22* to *2011-02-24*. Here the substation fed on average 0.167 and 0.122 megawatts per hour respectively into the power grid.

The weather factors corresponding to this effective power production can be seen in Figure 1 (c). Here, the temperatures were on average $-4.60$ and $-2.11$°C respectively. Figure 1 (d) reveals that the global solar radiation, measured in Joules/cm$^2$, averaged 5.154 in December and 4.483 in February. This leads to the hypothesis that photovoltaic power plants work most efficiently in temperature ranges between $-5$°C and $-1$°C and lose efficiency in temperature ranges above and below, even if the sun duration and solar radiation is high.

## 6 Conclusion and Future Work

We proposed a matrix-oriented approach to time series analysis. The central idea is the representation of all time series intervals with user-chosen statistical measures. Time series data can be compared and correlated across different parameters (by small multiples) and intervals (by the mapping to a matrix). We applied our idea to a comprehensive data set to increase understanding of the impact of weather parameters on the effectiveness of power generation. Our approach allows the visual correlation of thousands of time series with millions of measurement values. We plan to augment our approach with additional semantic zoom capabilities, showing time series charts for chosen matrix values at the lowest zoom level. Furthermore, we plan to incorporate a drag-and-drop window for an automatic correlation analysis of selected time series portions.

## References


[1] W. Aigner, S. Miksch, H. Schumann, and C. Tominski. *Visualization of Time-Oriented Data*. Human-Computer Interaction Series. Springer, 2011.

[2] N. Elmqvist, T.-N. Do, H. Goodell, N. Henry, and J.-D. Fekete. ZAME: Interactive Large-Scale Graph Visualization. *2008 IEEE Pacific Visualization Symposium*, pages 215–222, Mar. 2008.

[3] Energie Baden-Württemberg AG. Intelligent grid project (in german). `http://www.enbw.com/content/de/der_konzern/enbw_gesellschaften/regionalgesellschaft/aktuell/smartgrid/projekt_freiamt/index.jsp`. Online; accessed 26th June 2012.

[4] E. Keogh. A decade of progress in indexing and mining large time series databases. In *Proceedings of the 32nd international conference on Very large data bases*, pages 1268–1268. VLDB Endowment, 2006.

[5] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings of the 1996 IEEE Symposium on Visual Languages*, VL '96, pages 336–, Washington, DC, USA, 1996. IEEE Computer Society.

[6] T. Warren Liao. Clustering of time series data-a survey. *Pattern Recogn.*, 38(11):1857–1874, Nov. 2005.

[7] H. Ziegler, T. Nietzschmann, and D. A. Keim. Relevance driven visualization of financial performance measures. In *EuroVis*, pages 19–26, 2007.