TV-MV Analytics: A Visual Analytics Framework to Explore Time-varying Multivariate Data

Information Visualization
XX(X):1–19

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Aurea Soriano-Vargas[1], Bernd Hamann[2], and Maria Cristina F. Oliveira[3]

Abstract

We present an integrated interactive framework for the visual analysis of time-varying multivariate datasets. As part of our research, we performed in-depth studies concerning the applicability of visualization techniques to obtain valuable insights. We consolidated the considered analysis and visualization methods in one framework, called *TV-MV Analytics*. *TV-MV Analytics* effectively combines visualization and data mining algorithms providing the following capabilities: i) visual exploration of multivariate data at different temporal scales; and ii) a hierarchical small multiples visualization combined with interactive clustering and multidimensional projection to detect temporal relationships in the data. We demonstrate the value of our framework for specific scenarios, by studying three use cases that were validated and discussed with domain experts.

Keywords

Visual Analytics, Time-varying Multivariate Data, Visual Feature Selection, Data Visualization, Data Analytics

Introduction

Advances in software and hardware technology have expanded data collection and storage capacity to such an extent that ever increasing volumes of time-varying multivariate data are produced with limited effort. Data sets consisting of observations of multiple sequentially recorded time-stamped variables play an important role in many areas, e.g., environmental monitoring, social sciences, financial markets and population statistics. Typical time-varying multivariate data may describe heterogeneous variables, possibly with missing data values, recorded at high resolution over long time periods. In this context, data analysis can be extremely challenging, as it combines the need of understanding data evolution dynamics with difficulties introduced by high dimensionality (1).

Regardless of the domain, data analysis requires one to thoroughly explore the behavior of multiple variables and how they relate to observations, and identify how certain subspaces of variables are relevant for characterizing a particular behavior. There are scenarios where analysts start investigating a dataset with limited prior knowledge about the phenomenon to be understood. They may wish to explore the data considering different temporal scales, where scale refers to a certain temporal aggregation appropriate for analysis (e.g., minute, hour, days, month, year). A flexible context-preserving navigation strategy for data represented at multiple temporal aggregations is necessary (2).

Approaches in Visual Analytics (3), exploring the synergy between data mining from machine learning and data representations from information visualization, have been proposed to support processes for extracting information from data. It has been demonstrated that they can be successfully applied to various problem domains (4; 5; 6; 7). Fundamentally important is the interaction with graphical data representations, which supports integrating the human capability of recognizing visual patterns (inference capability) and a user's knowledge about the data analysis process. Visual analytics solutions assist users in creating adequate mental models of complex data, supporting the perception of global characteristics and identification of local behavior through user-driven exploration. This is necessary to confirm expected and

[1] Institute of Computing, University of Campinas, Brazil
[2] Department of Computer Science, University of California, Davis, U.S.A.
[3] Institute of Mathematics and Computer Sciences, University of São Paulo, Brazil

Corresponding author:
Aurea Soriano-Vargas, Institute of Computing, University of Campinas, Brazil.
Email: aurea.soriano@ic.unicamp.br

discover unexpected behavior (3). A visual analytics system should support users to validate hypotheses and perform exploratory tasks raising new questions about the data (8).

In a previous contribution (9), we introduced a solution for exploring ionospheric scintillation data. The solution is based on a time matrix representation and visualization, coordinated with similarity maps produced with multidimensional projection techniques. That solution is domain-specific, and it has great potential to be generalized and extended as a general framework for handling multivariate time data arising in diverse problem domains.

Inspired by our previous solution approach, we introduce a domain-independent visual framework in this paper, called *TV-MV Analytics*, to assist exploratory analytics for complex time-varying multivariate data sets. *TV-MV Analytics* is based on a general strategy guided by the objective of revealing global relationships in data variables and data instances. The framework offers the following capabilities:

(i) visual exploration techniques for inspecting multivariate values at different temporal scales;
(ii) a hierarchical small multiples visualization combined with interactive clustering and multidimensional projection algorithms, in support of recognizing temporal relationships.

These capabilities can be employed for interactive investigation of variable spaces as descriptors of the data instances and for assessing how different feature spaces characterize a certain behavior. We demonstrate the effective capabilities of *TV-MV Analytics* for data from three very diverse application domains: crime statistics from the State of São Paulo, Brazil; air quality data from an Italian city; and stock market data of U.S. companies. Results obtained were discussed with experts from the respective application domains.

We discuss related literature in Section **Related Work**, addressing two topics, namely exploratory visualization of time-varying multivariate data, and visual analysis of spaces of variables – including our previous approach to handle a database of measurements of ionospheric scintillation. The overall pipeline implemented is introduced and justified in Section ***TV-MV Analytics***. Section **Use Cases and Results** introduces the selected datasets and illustrates possible application of the *TV-MV Analytics* framework with three case studies illustrative of plausible data analysis scenarios on those datasets. We assess how it can assist in investigating a phenomenon and interpreting the behavior of its describing variables. Section **Conclusions and Future Work** summarizes our main results and contributions, and points to possible future research.

Related Work

Experts analyzing time-varying multivariate data face a growth in both the number of data instances and the number of data variables. This scenario poses multiple challenges to visualization designers (10), as standard visualizations, such as line graphs and variations, quickly become overcrowded and cannot be used. In this section we review research concerned with approaches for exploratory visualization of time-varying multivariate data, and visual representations to support the analysis of variable spaces.

Exploratory visualization of Time-Varying Multivariate Data

The simplest and oldest technique for visualizing time-varying data is the line graph (11), which is unfeasible for displaying many multivariate series. Alternative solutions to visualize long time-varying data were proposed, such as TimeSearcher visualization (12) to query on entity sets with one or more time-varying variables, and CircleView visualization (13) to represent multivariate data flows.

Many visualizations support **analysis of single or multiple temporal variables**, such as the Cluster and Calendar based Visualization (14) which presents the data patterns in a calendar and a 3D chart along the time axis, and VizTree (15) which aims to support pattern discovery in data without requiring prior knowledge. Turkay et al. (16) introduce two visualizations to assist the investigation and interpretation of structural changes in temporal clusters. A temporal clustering view conveys the structure and quality of a cluster set over time, which are assessed, regarding quality, with the Silhouette Coefficient (17) (employed for similar purposes in *TV-MV Analytics*); and a temporal signatures view summarizes statistical properties of the cluster structures over time.

Several contributions specifically target the **visualization of time-varying multivariate data**, to guide specialists in challenging tasks that require searching for relevant data patterns in bulky datasets. Correlated multiples (18) is a method based on small multiple visualizations (19) that places spatially coherent multiples views of chunks of time-varying data based on their dissimilarities. TimeSpiral (20) combines multiple views to assist users in analyzing and exploring periodic trends and correlations in time-varying multivariate data, supporting multiple time granularities. VIMTEX has been designed to assist geologists observing

temporal relationships in multivariate data describing concentrations of chemical compounds (5). It uses Parallel Coordinates for a time-varying view of the multivariate data, in combination with a density view to show univariate temporal distribution and a small multiples matrix view which shows bivariate correlations as time-series. Machado et al. (6) use pixel-based layouts called Player Attribute Heatmap (PAH) to visually summarize the events in a soccer match using as input the player positions over time.

Similarly to the solution introduced by Liu et al. (18), we adopt a small multiples visualization and a consistent placement strategy to display groups of related variables. Moreover, we consider arbitrary time periods and temporal scales by means of user-defined observation periods and data aggregations (20), which allows visual data explorations at multiple levels of detail.

Visualization and Analysis of Variable Spaces

There are many strategies for **user-driven exploration of datasets described in high-dimensional variable spaces**. Yang et al. (2003) (21) introduce the Visual Hierarchical Dimension Reduction (VHDR) method to generate lower-dimensional representation spaces exploring a hierarchy of the variables, displayed in a radial visualization called InterRing, which inspired our hierarchical visualizations. The Value and Relation (VaR) technique (22) aims to facilitate data exploration, visualization, and variable selection, helping users to understand relationships in variables and instances, where variables are projected with Multidimensional Scaling and represented by a glyph. The glyphs are "dense pixel displays" of the corresponding data values (similar to some of our strategies) mapped into a spiral arrangement of pixels.

Some approaches aim at **comparing multiple variable spaces** using estimators of their discrimination power, estimated from the uniformity of the histogram of the distances between data clusters obtained in the different spaces (23); or the concept of Relevance Feedback (RF), in which users, based on their knowledge, train the system informing the perceived relevance of the hits returned from a query (24). In *TV-MV Analytics*, we use the Silhouette Coefficient as a sub-space relevance estimator. The Dimension Projection Matrix/Tree (25) and the visualization model by Turkay (26) rely on projections for interactive visual analysis of multivariate data focusing on both the variables and the data instances. *TV-MV Analytics* also uses multidimensional projections to create 2D maps of variables or data instances to convey their similarity.

Visualization-assisted variable selection can also rely on **statistical techniques and measures**, such as the visual interface SmartStripes (27), a method for semi-automatic refinement of results from variable selection algorithms; the Hierarchical Clustering Explorer (HCE) interface (28; 29) that includes a variable ranking criterion and scatterplots and histograms to display color-coded ranking results; and the Rank-by-feature framework (30) that considers local variables selected in several data visualizations through the linking-and-brushing strategy. Turkay et al. (31) introduce the construction of representative factors to analyze structures in high-dimensional data, where the original variables and factors are combined into an interactive visual analysis cycle.

In (9), we introduced the concept of time matrix representation and its use in a small multiples visualization, in conjunction with similarity maps obtained via multidimensional projection. We demonstrated how these concepts could be combined in an integrated approach for exploring ionospheric scintillation data. The time matrices were constructed using pre-defined periods, i.e., explorations were performed considering static time units for analysis. For application in other scenarios, a generalization is required to allow for further flexibility in choosing the representation and visualization parameters. Time matrices were represented by means of feature vectors described by five standard statistical moments. However, other compositions of feature vectors may be employed to capture different data behaviors. We faced the problems of visualizing a large number of variables and loss of detail when visualizing data via the small multiples method. New strategies were required to support improved interactivity in the exploration. These challenges led us to re-think the interactive visualization software design, including generalized strategies. We also developed an interest in conducting studies with data from other domains to better understand the general requirements of time-varying multivariate data representation and exploration.

TV-MV Analytics Framework

TV-MV Analytics is a general framework that integrates multiple data mining and visualization techniques to assist in the exploration of multivariate time-varying data. It abstracts and generalizes common functions for processing time-varying multivariate data. Specifically, it supports i) data integration and transformation into a unified representation; ii) interpretation of large collections of multivariate temporal data; iii) understanding of incomplete data; and iv) analytical

reasoning about the underlying causal phenomena that generated the data.

We have identified our specific requirements following the corners of the "Data-Users-Tasks" design triangle (32). The requirements are:

- Data. The target data consists of large collections of multivariate, time-stamped records from certain domains, e.g., stock market, air quality, or crime indicator records. Data variables are quantitative and data instances can be aggregated for user-defined temporal intervals and organized into user-defined temporal cycles, e.g., hours, days, or months.
- Users. Users are domain experts or consultants in areas dealing with problems that require analysis of large collections of multivariate temporal data. It is expected they are familiar with standard data processing and mining techniques, such as interpolation, regression and correlation analysis and clustering algorithms.
- Tasks. The system should support the following tasks: (i) observing the behavior of variables over time, e.g., for finding patterns, trends, correlations, outliers, missing values etc.; (ii) exploring data behavior at multiple temporal scales by supporting aggregation of measured values over user-defined time intervals; (iii) comparing the behavior of multiple variables over user-defined time periods; (iv) identifying representative variables and representative sub-spaces of variables to explain observed behavior; and (v) predicting future values of variables based on historical data.

It utilizes a specific data representation, which in previous work has been found adequate to communicate properties of time-varying variables (9). Further, it supports inspecting the temporal behavior of variables, individually or as sub-space descriptors, at a user-defined temporal scale. A small multiples visualization component that conveys the temporal behavior of each variable is used in conjunction with clustering and multidimensional projection techniques, based on a consistent data representation. We define the input data handled by the framework, the derived time matrix representation and the analytics functions it provides. After an overview of the data handling pipeline, we describe the created visualizations supporting the pipeline.

Data representation

A time-varying multivariate dataset consists of time-stamped observations of n variables $\{\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3}, ..., \mathbf{x_n}\}$, recorded at a particular temporal scale (per minute, hour, day, month or year). Thus, each variable is described as:

$$\mathbf{x_i} = \{x_i^{t_1}, x_i^{t_2}, x_i^{t_3}, ..., x_i^{t_k}\} \tag{1}$$

From the recorded values, it is possible to derive new series by applying aggregation functions to the data values for a given time span, e.g., to obtain a maximum, minimum, average, median or standard deviation value. Let $x_i^{t_{se}}$ denote the result of applying to the values of variable $\mathbf{x_i}$ within time span $[t_s, t_e]$, an aggregation function $f : \mathbb{R}^l \rightarrow \mathbb{R}$:

$$x_i^{t_{se}} = f_{ag}(x_i^{t_s}, x_i^{t_{s+1}}, x_i^{t_{s+2}}, ..., x_i^{t_e}) \tag{2}$$

A data instance at time t_j given by p selected variables ($p \leq n$) can be represented as a vector $\mathbf{o^{(t_j)}}$ with p values:

$$\mathbf{o^{(t_j)}} = [x_1^{t_j}, x_2^{t_j}, x_3^{t_j}, ..., x_p^{t_j}] \tag{3}$$

where $x_i^{t_j}$ can denote an original or an aggregated value obtained with Eq. 2.

A multivariate time-varying data set is defined by time series describing multiple variables, see Eq. 4. In this matrix, each column corresponds to the time series relative to a particular variable, see Eq.1, and each row corresponds to a multivariate observation at a particular time stamp, i.e., a multidimensional data instance, see Eq. 3.

$$\mathbf{D} = \begin{bmatrix} x_1^{t_s} & x_2^{t_s} & x_3^{t_s} & \dots & x_p^{t_s} \\ x_1^{t_{s+1}} & x_2^{t_{s+1}} & x_3^{t_{s+1}} & \dots & x_p^{t_{s+1}} \\ x_1^{t_{s+2}} & x_2^{t_{s+2}} & x_3^{t_{s+2}} & \dots & x_p^{t_{s+2}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_1^{t_e} & x_2^{t_e} & x_3^{t_e} & \dots & x_p^{t_e} \end{bmatrix} \tag{4}$$

The flexibility of *TV-MV Analytics* in handling data from different domains is a result of how it handles the input time series, using the *time matrix representation* introduced in previous work (9). A time matrix represents the time series describing each variable, which records the temporal behavior of this particular variable for a defined observation period and a certain temporal scale. Variables, observation periods and temporal scales are user-defined.

Database user queries specify the target data for an analysis session, i.e., the time series relative to one or multiple selected variables that describe the sequence of data instances, see Eq. 4. A query must specify the initial and final dates d_s and d_e of the time period, the aggregation function f_{ag} and aggregation span t_{se}, see Eq. 2; it must also specify the target temporal unity cycle t_c. All input data referring to time is converted to milliseconds in a pre-processing stage.

The aggregation span t_{se} defines the temporal scale of the analysis, e.g., minutes, hours, months, etc.; values recorded

at a certain temporal resolution can be aggregated at a coarser resolution, e.g., by selecting average, maximum, etc. over the time span, as defined by f_{ag}. For each variable, the query retrieves a series of l data values, covering a cyclical time period. For example, a user may execute a query to retrieve data relative to 90 days, aggregated over hourly spans. In this case, $l = (90 \times 7.776e^9) \times (24 \times 3.6e^6)$ - in milliseconds-, and the temporal cycle t_c corresponds to the 24 hours of a day. Data values are linearly normalized to the range [0,1] to cancel the effects of the different orders of magnitude of means and variances, and possibly, interpolated, to reduce the effects of missing data.

The time matrix representation is illustrated in Figure 1. Each time series is organized as a matrix $\mathbf{M}_{n\times m}$, where an entry records a data value for this variable at a particular time, $n = (d_s - d_e)/t_c$, and $m = \lceil t_c/t_{se} \rceil$. Each row consists of m time-stamped values relative to a temporal cycle, and each column depicts the values, over all rows (temporal cycles), relative to a single time stamp. Algorithm 1 describes the procedure for computing the time matrix representation of a particular variable, from a series of retrieved values. The time matrix representation supports multiple analytics functions and visualizations in *TV-MV Analytics*, as discussed next.

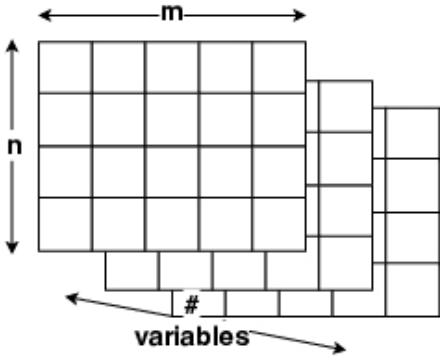


Figure 1. A time matrix is created per variable. Each matrix row depicts the m time-stamped values relative to a target temporal cycle and each column depicts the values relative to the same time-stamp, over all temporal cycles.

Data Analytics Functions

While multivariate data analysis approaches usually operate directly on given data, this approach is not feasible for time-varying multivariate data because a) there are too many data instances; b) it is important to consider the temporal sequence of data; and c) there is a need to explore the behavior of the multiple variables when their relationships become increasingly complex.

The derived time matrix representation is highly informative concerning the temporal behavior of individual

input : $\mathbf{x_i} \Rightarrow$ time series values of i^{th} variable
$d_s \Rightarrow$ initial date
$d_e \Rightarrow$ final date
$t_{se} \Rightarrow$ aggregation time span
$t_c \Rightarrow$ temporal cycle
$f_{ag} \Rightarrow$ aggregation function
output: $\mathbf{M_{i_{n\times m}}} \Rightarrow$ time matrix of i^{th} variable
$n = (d_s - de)/t_c$
$m = \lceil t_c/t_{se} \rceil$

for $a \leftarrow 1$ ***to*** n **do**
  for $b \leftarrow 1$ ***to*** m **do**
    /* index position */
    index = a *m + b;
    /* initial date */
    $d_s^{'} = ds + index * t_{se}$;
    /* final date */
    $d_e^{'} = d_s^{'} + t_{se}$;
    /* subset of values of $\mathbf{x_i}$ */
    $\mathbf{x_i^{'}} = \mathbf{x_i}\left[d_s^{'} \rightarrow d_e^{'}\right]$;
    $m[a,b] = f_{ag}(\mathbf{x_i^{'}})$;
  end
end

Algorithm 1: Time Matrix Construction for variable $\mathbf{x_i}$

variables, and flexible in that it supports inspecting data at different temporal scales. In order to further enhance the analytical capabilities of the framework, we derive representative feature vectors from the time matrices, for use in subsequent mining and visualization tasks. In order to satisfy task requirement (i) the features must capture the relevant characteristics of the variable behavior over time.

A feature extraction function is used to obtain feature vectors $\mathbf{V_i}$ from the time matrix representation $\mathbf{M_i}$, as defined in Equation 5:

$$\mathbf{V_i} = f_{mom}(\mathbf{M_i}). \quad (5)$$

TV-MV Analytics supports multiple feature extraction functions f_{mom} to generate feature vectors representative of the time matrices. For example, statistical moments (color moments) (33) and Abo-Zaid moments (34) can be computed.

In order to provide solutions in support of task requirements (iii) and (iv), feature vectors can be clustered as descriptors of time matrices to obtain groups of variables with globally similar behavior over the observation period captured in a time matrix. The clustering operation is defined as a function f_{clus} that uses a distance function f_{dis} for the set $\mathbf{V}$ of feature vectors and returns a vector of cluster labels $\mathbf{L}$ for the set $\mathbf{V}$, see Eq. 6. The system currently supports Euclidean distance and Inverse Pearson Correlation as dissimilarity functions.

$$\mathbf{L} = f_{clus}(\mathbf{V}, f_{dis}(\mathbf{V})) \quad (6)$$

The framework offers choices for selecting a clustering algorithm, including K-means and X-means clustering choices that are popular and are used in a variety of application domains (35). Further, it supports a variation where the optimal number of clusters k is set to the value that yields the best Silhouette Coefficient (SC). SC values are in the range $[-1, +1]$, with higher values indicating better cluster cohesion and separability. Users are informed about the SC value of any cluster model, and they can interact with multiple visualizations, see sub-section *Visualizing the Behavior of Data Variables*; they can also modify the clustering according to their perception of element similarity, causing SC values to be updated accordingly. The categorical colormaps available at ColorBrewer* (36) are employed to differentiate the clusters in the visualizations.

For subspace analysis and variable selection purposes, (according to task requirement (iv)), the framework offers five possibilities to identify representative sub-spaces, illustrated in Fig. 2: (a) the set of variables defined by the cluster medoids is initially assumed to be the default set of representatives as a descriptor of the data instances; (b) analysts can modify the default set selecting a different variable as cluster representative; (c) analysts can select multiple representative variables per cluster, picking those with smallest distances to the virtual centroid; (d) analysts may use linear regression to select the best correlated variable; or (e) they can utilize multiple linear regression to remove variables that do not contribute to the regression function.

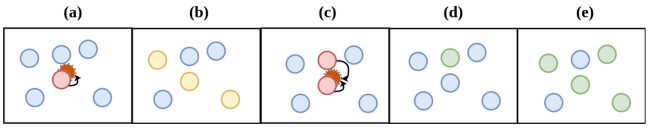


Figure 2. Five strategies for variable sub-space selection. Orange circles represent virtual cluster centroids; red circles are points closest to virtual centroids; yellow circles represent user selections; green circles represent results from linear regression.

TV-MV Analytics embeds a decision tree and multilayer perceptron classification algorithms, as well as linear regression and simple linear regression algorithms. Analysts can execute them to obtain further information regarding alternative sub-spaces of variables considered and assess potential data descriptors effective for characterizing a target variable, as defined in Eq. 7.

$$R_{clas/reg} = f_{clas/reg}(\mathbf{V^n}) \quad (7)$$

where $\mathbf{V^n}$ is a feature vector defined by a subset of representative variables. Decision trees and neural networks were chosen since they are widely known and have been successfully applied to pattern classification problems (37).

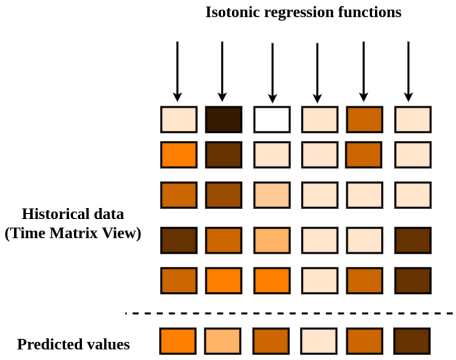


Figure 3. Prediction task using Isotonic Regression.

Moreover, the time matrix representation supports prediction tasks based on historical data, according to task requirement (v). A functionality is included to predict the series of values in a row from a given time matrix, from the values recorded in previous rows. This is achieved by applying Isotonic Regression functions to the time matrix columns, as illustrated in Fig. 3. The Isotonic Regression function f_{iso} can be understood as a least squares approach performed under order restriction. It is defined as a monotonic function best fitting the original data instances, minimizing Eq. 8

$$f_{iso}(\mathbf{T}, \mathbf{M}_{ib}) = \min \sum_{j=1}^{n} w_j (M_{ibj} - T_j)^2, \quad (8)$$

where $\mathbf{M}_{ib}$ refers to column b of time matrix $\mathbf{M}_i$, $\mathbf{T}$ is the time stamp (in millisecond) of element m_{ibj}, and w_j is set to 1. The result of the Isotonic Regression function is a model that fits the isotonic function of the explanatory variables to estimate the response variable. This model is employed to obtain a future time instant T_{n+1} (in milliseconds) associated with the value to be predicted, p_b, as defined by Eq. 9:

$$p_b = predict(f_{iso}(\mathbf{T}, \mathbf{M}_{ib}), T_{n+1}). \quad (9)$$

If the actual data at this time stamp is known, the correlation between the predicted and actual values is computed and displayed. Thus, an analyst may assess and compare the prediction capability of the model, considering data from distinct temporal ranges.

*http://colorbrewer2.org/

Data Exploration Pipeline

Fig. 4 shows all stages of a general data exploration process. For a given dataset a user first issues a query to retrieve a subset of the data for analysis. S/he may select a subset of variables, define the time period of the observations, the temporal scale and the data aggregation function (Eqs. 2, 3. The query returns a matrix of data instances for the specified time period, according to Eq. 4. Each variable is given by a series of time-stamped values, possibly aggregated over a specified time window, as in Eq. 2. The aggregation defines the temporal scale of the analysis. Variable ranges are linearly normalized to the range [0,1] to cancel the effects of the different orders of magnitude of means and variances. When data values are missing, a user may choose to preserve them as missing or replace them by estimated values obtained by interpolation.

The time matrices representing each variable are constructed, following Algorithm 1, and representative feature vectors, see Eq. 5, are derived from them for subsequent mining and visualization. An initial clustering of the feature vectors, see Eq. 6, is computed to convey groups of variables with similar behavior over the selected time period.

Finally, visualizations of both variables and instances can be created, detailed in the remainder of this section. The visualizations of variables include a hierarchical similarity map, correlation matrices, hierarchical time matrices, and a time-circular diagram, conveying further insight into temporal behavior and correlations. There are also normalized line-graphs views and aggregated and hierarchical similarity-based visualizations of the instances as described by user-defined sub-spaces of variables. Interaction with these visualizations, combined with results from data analytics functions for classification, regression and prediction, allow experts to assess the role of alternative sub-spaces of variables as representative data descriptors. The multiple time matrix and similarity based visualizations are coordinated for enhanced user interaction.

Visualizing the Behavior of Data Variables

TV-MV Analytics offers four visualizations for depicting variable behavior over time, driven by the analytical tasks relevant for multivariate temporal datasets.

Hierarchical Time Matrix Visualization The time matrix view introduced by Soriano et al. (9) summarizes temporal behavior of a particular variable over a target period. Each time matrix representation M (see Alg. 1) is visualized as a rectangular area that includes a header area with information about the variable and a main area split into cells representing its corresponding time matrix entries, see Figures 6 and 8. The header areas are colored to reflect the cluster of the corresponding variable. Cells depicting matrix entries are colored using the heated-object colormap (38), with brighter colors representing lower values and darker colors representing higher values. Cells with red borders indicate that the corresponding entry consists of interpolated data, and gray cells represent missing values.

Time matrix views can have substantial space needs, which is aggravated when the matrices describe a long, high-resolution time period, thus requiring larger visualization areas. This poses a disadvantage for scenarios where many variables must be analyzed. It is difficult to display the small multiples, and it is hard for an analyst to inspect many time matrices simultaneously. Hence, we have extended that representation to a hierarchical version of the small multiples view, in order to provide an improved solution for task requirements (i), (ii), (iii), and (iv). The hierarchical time matrices view relies on the cluster tree resulting from the X-means algorithm.

Leaf nodes represent single variables, whereas nodes higher up in the hierarchy summarize the content of a cluster by showing the time matrix of its medoid element. Each cluster is sub-clustered using X-means (see Eq. 6), defining new sub-cluster nodes displayed as their corresponding medoids, done repetitively until reaching leaf nodes. An analyst can explore different tree levels, exploring sub-clusters represented by their medoids. The hierarchical tree representation is coordinated with the hierarchical similarity map, as illustrated in Fig. 8 where the user selected a variable in the map, which is highlighted in cyan both in the map and in the small multiples time matrix visualization.

Time Circular Diagram The time matrices are capable of conveying variables' global behaviors while preserving detail. However, analysts may be interested in inspecting whether a certain variable displays correlations in time intervals within the observed period, thus motivating the introduction of the time circular diagram visualization, to provide additional mechanisms addressing task requirements (i), (ii), (iii) and (iv). This visualization is also split into two areas (as shown in Fig. 13), a header area with information about the variable and a main area.

The main area displays the normalized entries of a time matrix $\mathbf{M}$ (obtained with Alg. 1), but cells are now arranged in a circular, clockwise distribution, again adopting the heated-object colormap (38). An analyst can define a target

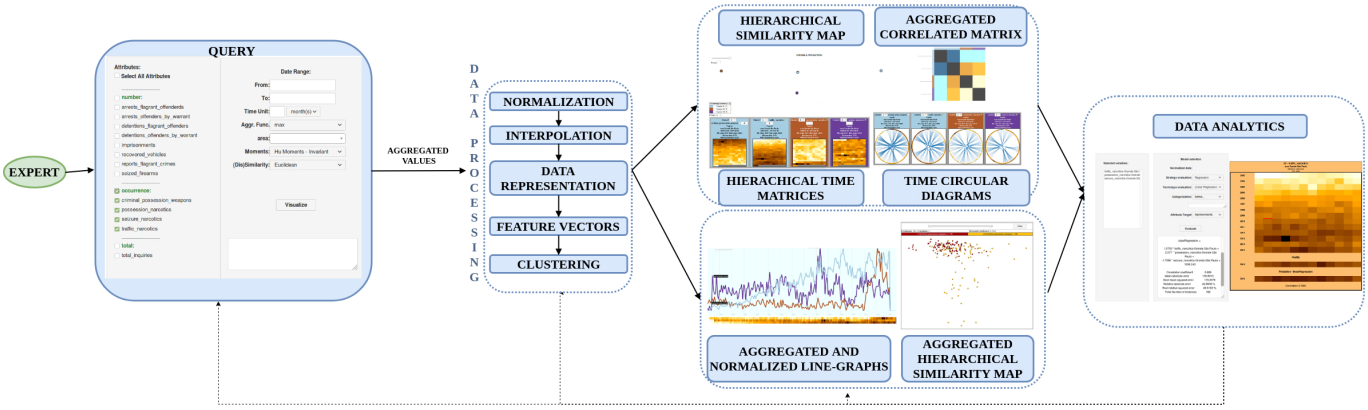


Figure 4. Stages of data exploration. A user query retrieves aggregated instances described by a subset of variables over a time range. Values of each variable (normalized) are represented in a time matrix, where each entry stores a unit value. Feature vectors describing each time matrix are clustered into groups of variables with similar temporal behavior, employed in visualizations of sub-spaces of variables as well as data instances.

temporal window to assess possible correlations within sub-periods the observation period. Lines are used to connect sub-periods when their correlation is above a user-defined threshold, applying an edge bundling strategy (39) to reduce clutter and group similar edges. The sub-periods can be reduced by using the Haar wavelet (40) to simplify the time series representation and produce an approximation at a lowest resolution level. The maximal resolution $level$ is calculated using the sub-period size sp_size, defined by Eq. 10. This feature is useful for exploring data relative to very long temporal intervals (as shown in Fig. 16).

$$\begin{aligned} sp\_size &= 2^{level} \\ level &= \log_2 sp\_size \end{aligned} \tag{10}$$

Hierarchical Similarity Map of Variables A similarity map depicts a 2D space embedding of the m-dimensional feature vectors (see Eq. 5) describing the variables' time matrices. In this representation, the pairwise point distances are taken as proxies of the point distances in the original m-dimensional space (41). Therefore, variables with *more similar* temporal behavior (as captured by the feature vector of their corresponding time matrix) are relatively closer in the 2D map, whereas those with *dissimilar* behavior are farther apart, generally. The similarity maps of variables assists users in assessing and modifying the clustering model and identifying representative variables when looking for representative sub-spaces of variables as effective data descriptors, in order to address task requirements (i), (iii), and (iv).

In *TV-MV Analytics* the 2D embedding is computed with the IDMAP multidimensional projection (41), taking the pairwise Euclidean distances or the Inverse Pearson correlations as proxies for pairwise dissimilarity of feature vectors. The map is shown as a point plot in which each circle depicts a variable and is colored to convey its associated cluster, preserving the same color coding adopted in the small multiples matrix view.

The map conveys overall similarity between variables and their implicit groupings, regardless of their explicit cluster assignments. The map can help analysts in identifying *natural* group neighborhoods. This is the reason for showing the map view in connection and coordinated with the small multiples matrix view. Inspecting both in combination provides an analyst with insights needed to decide whether the explicit cluster assignments should be modified and alternative cluster models should be investigated or not.

In order to reduce the undesirable effect of overlapping circles, the map can be explored at multiple levels of detail. We implemented a KD-tree-based spatial representation (42) of the points embedded in 2D space. The root represents the set of all projected variables, whereas the the leaf nodes represent individual variables. At higher tree levels the map shows bigger circles that represent multiple variables. The map is initially displayed at the leaf level, but a user can navigate up the tree hierarchy to reduce the level of detail, and then down to increase it. Circles that represent multiple variables assigned to a single cluster are rendered with the cluster color, while circles rendered in gray represent a heterogeneous group of variables that have been assigned to multiple clusters. Hovering over a circle prompts the display of more detailed information about the variables it represents.

Variable Correlation Matrix A correlation matrix of time matrices is computed for all variables, or for a user-selected subset, see Fig. 11. This provides a complementary view

to verify cluster cohesion to efficiently identify groups of variables with similar temporal behavior, as identified by task requirements (i), (ii), and (iii). Pairwise correlations between variables are computed using Pearson's correlation coefficient, as implied directly by their corresponding data values, rather than their feature vectors.

Variable correlation is displayed as a heatmap visualization, where each cell depicts an entry of the correlation matrix. Values are mapped to a color scale where darker blue shades indicate strong negative correlations and darker brown shades indicate high positive correlations. Matrix rows and columns are sorted based on the clustering of the variables, and clusters are signaled by the colored borders at the top and the left side. Analysts can inspect it to assess the similarity of variables assigned to a single cluster.

Normalized time series line graphs of the variables can also be inspected, with lines colored according to the variable's cluster. Motivated by task requirements (i), (ii), and (iii) and inspired by the temporal pixel-oriented layout introduced by Machado et al. (6) this view is associated with an array showing the corresponding matrix values, as in a timeline, as illustrated in Fig. 7. These visualizations are coordinated, e.g, in the example, a data point has been selected in the pixel-based layout, highlighted with a border in cyan, which is also highlighted in the corresponding line graph. These combined and coordinated views facilitate identifying common patterns and values in different time windows.

Visualizing the behavior of data instances

It may be also relevant to visualize the behavior of data instances, in particular, 2D similarity maps can convey groups of similar data instances. Multiple maps can be created considering the data instances described by alternative sub-spaces of variables (see Eq. 3), allowing one to compare the behavior of distinct descriptor spaces in relation to a target variable, in support of task requirement (iv).

After some empirical investigation we chose the Least Square Projection (LSP), introduced by Paulovich et al. (43) for this purpose. The technique attempts to preserve local neighborhoods identified in the original data space. Initially, it projects a subset of so-called control points with a dimension reduction method that preserves distance relationships accurately. It constructs a linear system of equations considering point neighborhoods in the original space and the projected coordinates of the control points. The solution of this linear system defines the remaining positions in the low-dimensional space.

Each data instance is shown as a colored circle, mapping either a category or a scalar variation of a selected target variable. Pairwise distances can be computed with the Euclidean distance or the Inverse Pearson correlation. Analysts can filter out data instances based on the value of the target variable, which is useful to reduce overlapping and conduct more focused data analysis.

Use Cases and Results

The *TV-MV Analytics* framework is applicable to domains where large time-varying multivariate data arise and must be understood. It can be used for interactive exploratory data visualization considering different temporal ranges and scales, from a variable and/or instance perspective. We have used three time-varying multivariate datasets from distinct domains, namely crime statistics of the state of São Paulo, Brazil, air quality measurements from an Italian city, and U.S. stock market data, summarized in Table 1.

In the case studies we considered several realistic questions and scenarios to exercise multiple exploration paths and illustrate the framework's potential, as described in the remainder of this Section.

In the case studies we considered several realistic questions and scenarios to exercise multiple exploration paths and illustrate the framework's potential, as described in the remainder of this Section.

Table 1. Datasets analyzed in the case studies.

Domain	# Var.	Rate	# Instances	Interval
Crime Statistics	14	month	2,366	Jan 01, 2002 Feb 01, 2017
Air Quality	13	hour	9,357	Mar 10, 2004 Apr 04, 2005
Stock Market	7	day	1,773,230	Jan 01, 1995 Nov 30, 2016

Crime Statistics of São Paulo State, Brazil

Crime statistics provide information about public safety and relevant for planning police actions and investments. In the state of São Paulo, the Secretariat of Public Security is responsible for data collection and analysis to support crime prevention and repression. A public crime statistics dataset is available[†] storing several indicators of criminal occurrences in 12 mesoregions, recorded monthly from January 1, 2002 to February 1, 2017.

The data instances comprise 13 variables recording the numbers of monthly occurrences related to the following: (1)

[†]Secretariat of Public Security, State of São Paulo, Brazil, Police Productivity http://www.ssp.sp.gov.br/Estatistica/Pesquisa.aspx, (December 28, 2017)

possession of narcotics; (2) traffic of narcotics; (3) seizure of narcotics; (4) criminal possession of weapons; (5) seized firearms; (6) official reports of flagrant crimes; (7) detentions of flagrant offenders; (8) detentions of offenders by warrant; (9) arrests of flagrant offenders; (10) arrests of offenders by warrant; (11) imprisonments; (12) recovered vehicles; and (13) total inquiries. We added variable (14) informing the homicide rates over the period, also reported by the state Secretariat of Public Security.

The data then available was investigated in an earlier study by Arvate and Souza (44), who estimated the impact on crime indicators of a law approved to regulate firearm adoption by municipal police forces in 2003. They constructed indicators for each state region for years 2002, 2004, 2006, 2009 and 2012, using the following indicators: total number of inquiries; individuals arrested in flagrant; drug seizure; recovered vehicles; prisons issued and homicide. They created regression functions using these indicators as dependent variables and concluded that armed municipal police forces contributed to a reduction in crime indicators after 2003. They also attributed to this factor a significant increment observed in the number of individuals arrested in flagrante.

We discussed the value of our analysis framework for this scenario with a police officer. He stated that the crime indicators explored provided a meaningful and helpful way to assess criminal activity and police effectiveness in the different regions. Indeed, some indicators, such as number of seized firearms, occurrences of narcotics traffic, and number of imprisonments, can be viewed as indicators of police activity. Hence, an increase in their occurrence can be interpreted as a positive indicator of stronger police action. Our framework was tested by the police expert, who interacted with it for about an hour. It took no more than 10 minutes for him to utilize the tool properly and begin to recognize important behavior.

Our case study considered two illustrative data analysis questions: **(Q1)** Can we confirm that fire-armed police impact crime rates?, and **(Q2)** How do crime indicators relate to homicide rates? Analysis considered variable dissimilarity measured with the Inverse Pearson Correlation and the Abo-Zaid moments as time matrix features, emphasizing data variability.

Our exploratory investigation covers a longer period than the original study, from January 31, 2002 to February 28, 2017 (with time matrices shown in Fig. 6), in order to determine whether the conclusion still holds. We considered the same six indicators that had been analyzed in a previous study (44): *total inquiries*, *arrests of flagrant offenders*, *seizure of narcotics*, *recovered vehicles*, *imprisonments*, and *homicide rates*. The target data was retrieved with the query operation as illustrated in Fig. 5. It retrieves the observations relative to the six target indicators as stored in the database, without the application of an aggregation function; values were normalized. Algorithm 1 was executed to create the time matrices from the resulting set of time series. Feature vectors with Abo-Zaid moments as features, see Eq. 5, were extracted from the time matrices, and an initial clustering was generated.

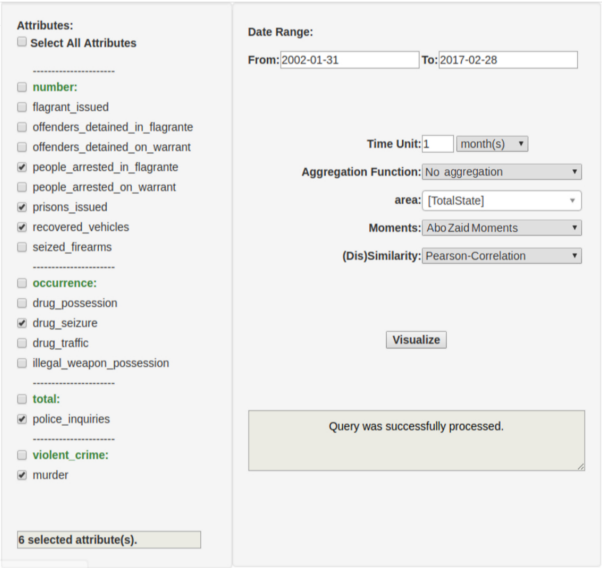


Figure 5. Query operation to retrieve six indicators: *total inquiries*, *arrests of flagrant offenders*, *seizure of narcotics*, *recovered vehicles*, *imprisonments* and *homicide rates*, covering a longer period, from January 31, 2002 to February 28, 2017.

Three clusters were identified ($SC = 0.44$) using Inverse Pearson Correlation as dissimilarity measure and the SC-optimized K-means clustering algorithm. The small multiples time matrix visualization was created, and a color mapping selected using distinct colors for each cluster. A blue cluster (C1) includes the indicators concerning *arrests of flagrant offenders* and *seizure of narcotics*; a yellow cluster (C2) includes *imprisonments* and *homicide rates*; and a pink cluster (C3) includes *recovered vehicles* and *total inquiries*. We observe a reduction in all indicators for a small period after 2003. A significant increase is observed from 2011, except for the indicators *homicide rates* and *recovered vehicles*, which exhibit different behavior. The numbers of *recovered vehicles* decreased in 2007 and 2008, but it increased later. The *homicide rates* decreased after 2004.

As clusters were computed using the inverse of Pearson Correlation as dissimilarity measure, we found that cluster C2 conveys the negative correlation of the indicators

Figure 6. Time Matrices of indicators *total inquiries*, *arrests of flagrant offenders*, *seizure of narcotics*, *recovered vehicles*, *imprisonments*, and *homicide rates* (January 31, 2002 to February 28, 2017).

imprisonments and *homicide rates*, i.e., the number of reported *homicide rates* decreased as the number of *imprisonments* increased. This is noticeable in the patterns of both time matrices in the yellow cluster. This observation supports the alleged positive impact of fire-armed municipal police, of short duration. Other factors could be causing an increase in the crime indicators reported. This reasoning is supported by the time series graphs depicting the behavior of these indicators, see Fig. 7, where reported numbers increase after 2010.

As a second investigation, we wanted to verify the role of the remaining 13 indicators for characterizing the behavior of the *homicide rates* indicator. In this case, the data was retrieved by sampling maximum values recorded over one-month periods for each year from January 2002 to February 2017, and values were normalized. Algorithm 1 was executed to create the time matrices relative to the resulting set of aggregated time series, and feature vectors defined by the Abo-Zaid moments, see Eq. 5, were extracted from the time matrices.

We initially inspected the small multiples time matrices view combined with the IDMAP similarity map for the 14 indicators for the State of São Paulo, see Fig. 8. We clustered feature vectors with the SC-optimized K-means algorithm and the Inverse Pearson Correlation as dissimilarity measure, yielding four clusters, with $SC = 0.51$, indicating a high-quality clustering. The views of the time matrices and similarity map confirm that variables in the same cluster are highly correlated regarding their temporal distribution of values. For instance, in cluster C1 (blue) we notice that the three indicators *official reports of flagrant crimes*, *imprisonments* and *arrests of flagrant offenders* are highly correlated. The indicators in clusters C2 (pink) and C3 (red) show lower correlation, confirmed by the similarity map. Cluster C4 (yellow) merges two groups with a strong positive correlation, a first group containing *criminal possession of weapons*, *seized firearms* and *homicide rates*, and a second group containing *seizure of narcotics* and *traffic of narcotics*. These two groups present high negative correlation. The selected medoids are: *official reports of flagrant crimes*; *detentions of offenders by warrant*; *arrests of offenders by warrant*; and *seized firearms*. Considering this, cluster C4 was split into two subgroups. The adjusted model, shown in Fig. 9, is defined by five clusters and has an improved value of $SC = 0.56$. A new medoid, *seizure of narcotics*, has been added to the representative indicators, see Eq. 7.

Visually, the indicators most correlated with *homicide rates* are *criminal possession of weapons* and *seized firearms*. In order to further investigate this issue, we compared the results of linear regressors, see Eq. 7, for the set of data instances described by alternative sub-spaces of indicators, see Fig. 2, with indicator *homicide* as target. The following sub-spaces were considered as descriptors:

1. **All indicators:** 13 indicators, excluding *homicide rates*
2. **Four cluster medoids (from initial automatic clustering):** *official reports of flagrant crimes*; *detentions of offenders by warrant*; *arrests of offenders by warrant*; and *seized firearms*
3. **Five cluster medoids (from user-adjusted clustering):** *official reports of flagrant crimes*; *detentions of offenders by warrant*; *arrests of offenders by warrant*; *seized firearms*; and *seizure of narcotics*
4. **Nine indicators from automatic linear regression:** *traffic of narcotics*; *criminal possession of weapons*; *seized firearms*; *official reports of flagrant crimes*; *detentions of flagrant offenders*; *arrests of flagrant offenders*; *arrests of offenders by warrant*; *imprisonments*; and *recovered vehicles*
5. **One indicator obtained by simple linear regression:** *criminal possession of weapons*
6. **Five indicators considered by Arvatez et al. (44):** *arrests of flagrant offenders*; *seizure of narcotics*; *imprisonments*; *recovered vehicles*; and *total inquiries*

Results are summarized in the projection maps shown in Fig. 10 (except in (e)) in which the circle colors map the indicator *homicide rates*, accompanied by the regression

Figure 7. Time series graphs of *arrests of flagrant offenders*; *seizure of narcotics*; *imprisonments*; *homicide rates*; *recovered vehicles*; and *total inquiries* (January 31, 2002 to February 28, 2017). Most indicators show an increase after 2010.

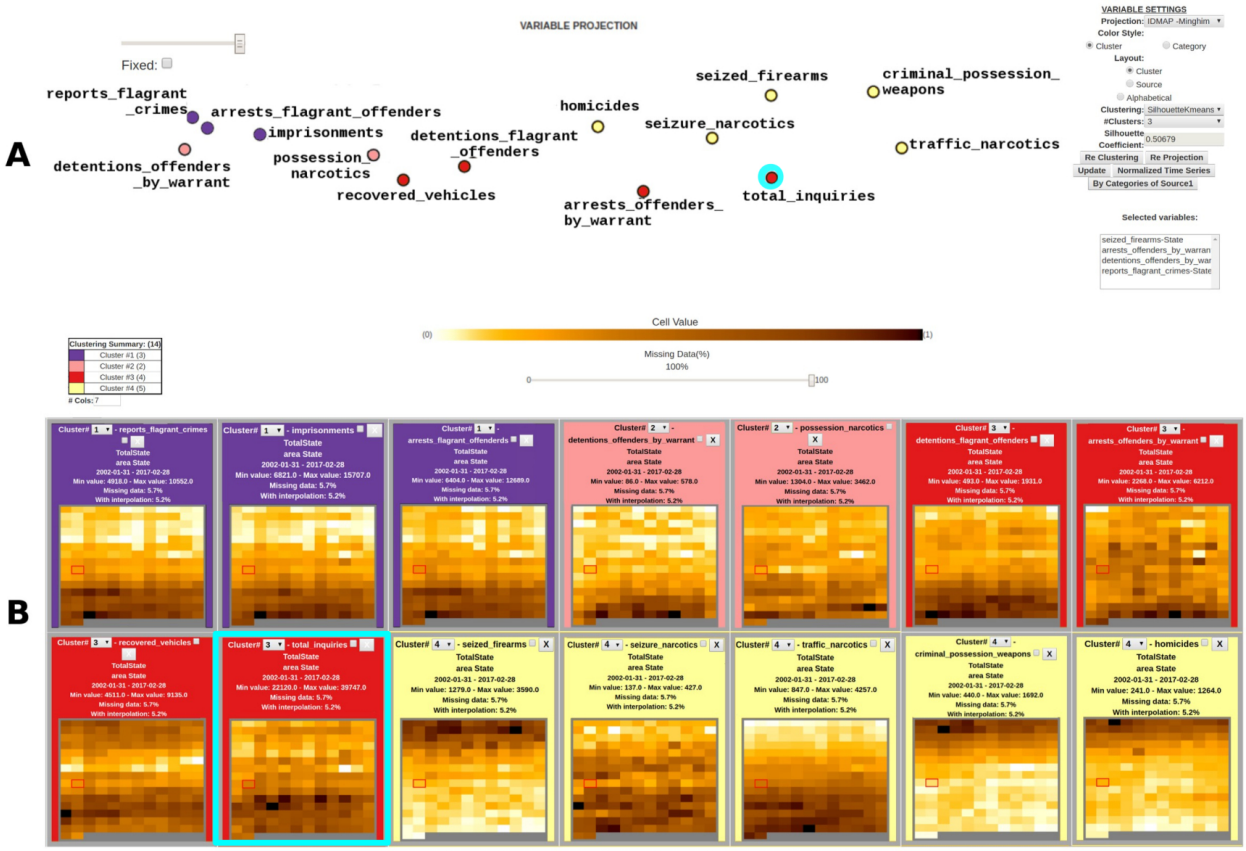


Figure 8. Crime indicators São Paulo state, Brazil): (A) IDMAP similarity map view, with each circle depicting an indicator; and (B) small multiple time matrix view, where matrix entries show the maximum values recorded over a monthly period, for each year. Four clusters are identified.

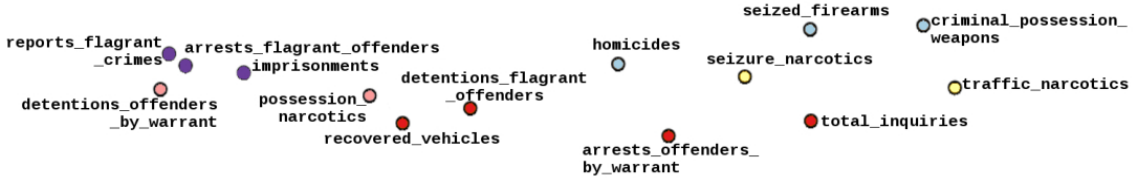
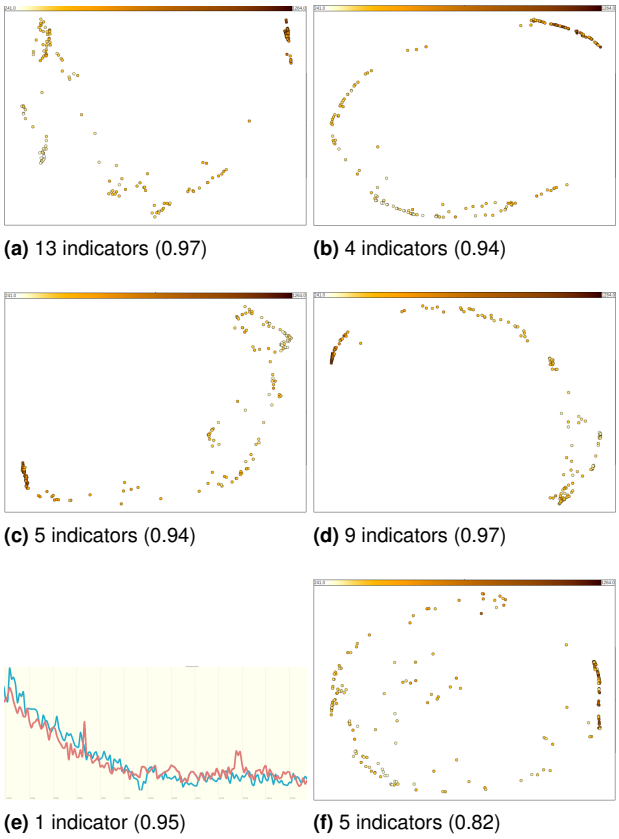


Figure 9. IDMAP similarity map view of the 14 indicators depicted in Fig. 8, now considering a user-adjusted cluster model (five clusters).

correlation indices between the target variable and the variable sub-spaces. They suggest that multiple variable sub-spaces can properly characterize this indicator, and good prediction models can be attained. Notice that in most similarity maps high homicide rates (represented by dark circles) are placed distant from the low homicide rates

(light colored circles). This analysis confirms that the visual interaction is beneficial for identifying the best subset of indicators to characterize the indicator *homicide rates*. Fig. 10(e) shows the time series of variables *criminal possession of weapons* (in blue) and *homicide rates* (in red), which also show strong correlation.



(a) 13 indicators (0.97) **(b)** 4 indicators (0.94)

(c) 5 indicators (0.94) **(d)** 9 indicators (0.97)

(e) 1 indicator (0.95) **(f)** 5 indicators (0.82)

Figure 10. Similarity maps and regression correlation of data instances described by the distinct indicator subspaces, which suggest that different variable subspaces represent data behavior well (numbers correspond to the correlation coefficients).

The domain expert concluded that the *TV-MV Analytics* made it possible to study relevant correlations, inconsistencies, and behavior over time, regarding both crime indicators and police activity. She affirmed that the framework allowed her to identify regions with high concentrations of certain types of crimes, suggesting the need for reinforcement of police activity.

Air Quality Data of an Italian City

The second dataset[‡] concerns urban environment pollution monitoring, with measurements obtained from four sensing devices deployed in an Italian city with heavy car traffic, from March 10, 2004 to April 04, 2005. One of the devices is referred to as the MOX sensor, as it is equipped with five Metal Oxides (MOX) sensors (45), and their concentrations were recorded hourly concerning our dataset. The other sensing device is a conventional air pollution monitoring station (ST) equipped with spectrometer analyzers to provide reference data. Two commercial temperature and humidity sensors were co-located with the pollution monitoring sensors.

The following six pollutant gases were tracked (some of them by both devices): carbon monoxide CO (tracked by ST and MOX); non-methanic hydrocarbons $NMHC$ (ST and MOX); nitrogen oxide NO_x (ST and MOX); nitrogen dioxide NO_2 (ST and MOX); ozone O_3 (MOX); and benzene C_6H_6 (ST and MOX). The data instances are therefore described by 13 variables including the six pollutant gases information tracked by the respective sensors, plus the the three sensor information of temperature, relative humidity, and absolute humidity (hourly averages).

This data set was previously investigated by Vito et al. In the first study (45) the authors employed back-propagation neural networks to predict benzene (C_6H_6) concentration, with low estimation errors. They found that the NMHC pollutant has a notable benzene-related component. The second analysis (46) also used back-propagation neural networks to study the influence of variable selection and correlation analysis for estimating CO, NO_x and NO_2 concentrations. Interestingly, the best performance values for estimating NO_2 were obtained when using data from all sensors. Results concerning CO estimation revealed that $NMHC$ can follow the concentration of CO, and that estimations deteriorate when NO_2 and NO_x are added.

We discussed this scenario with an environmental engineer. She stated that the pollution problem in urban environments is highly affected by the dense distribution of outdoor air pollutants (some of them covered in this section) due to city design.

According to the specialist, the correlation study may help to identify the cause of pollution, e.g., pollutants emitted by the same sources or produced by the transformations induced by chemical mechanisms. In this context, *TV-MV Analytics* supports further exploration of a correlation matrix visualization of selected variables. We created Correlation Matrix views to verify the variable correlations considering the original hourly values recorded, which were retrieved as stored in the database, see Fig. 11.a; and considering the daily maximum values recorded, which were retrieved from the database with a query setting the maximum value over a 24 hour period as the aggregation function (as defined

[‡]UCI, Air Quality Data Set, https://archive.ics.uci.edu/ml/datasets/Air+Quality, (December 28, 2017)

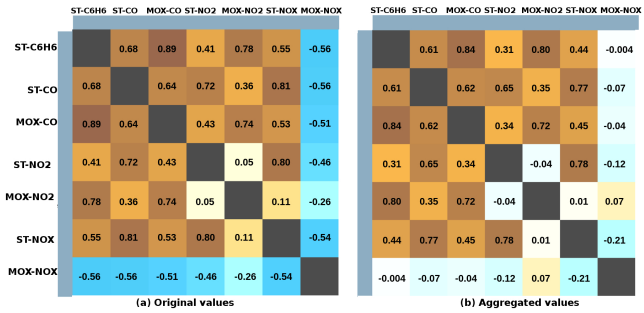
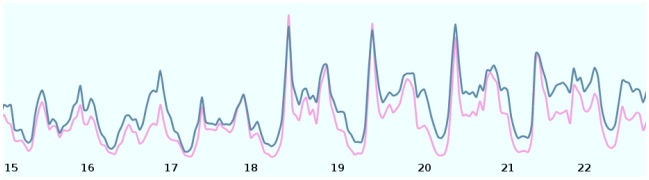
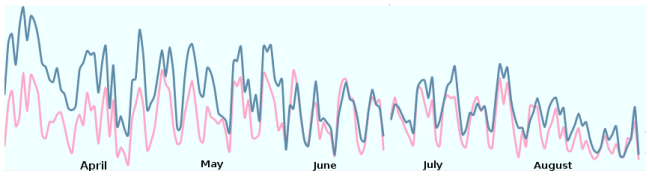


Figure 11. Correlation Matrix of original and aggregated values measured for the following variables: C_6H_6, plus CO, NO_2 and NO_x (reference and MOX), from March 10, 2004 to April 04, 2005. Dark blue shades indicate strong negative correlation and dark brown shades indicate high positive correlation.



(a) $ST\_C_6H_6$ (pink) and MOX_CO (blue) (original measurements, May 2004).



(b) $ST\_C_6H_6$ (pink) and MOX_CO (blue) (aggregated measurements, March-August 2004)

Figure 12. Line graphs depicting normalized series of measurements of $ST\_C_6H_6$ (pink) and MOX_CO (blue): (a) original hourly values (May 15 to May 22, 2004); and (b) aggregated daily values (March to August, 2004). Patterns can be observed at both temporal scales.

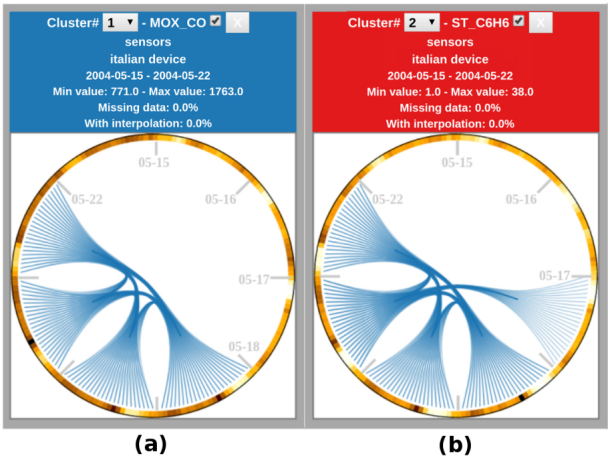


Figure 13. Time circular diagram of variables $ST\_C_6H_6$ and MOX_CO, considering values for the period May 15 to May 22, 2004, where values were grouped in one-day sub-periods.

in Eq. 2), see Fig. 11.b. Our goal was to assess how the data reduction due to aggregation impacts the outcome of a correlation analysis. After aggregation, the time series have been reduced from $9,384$ to 391 samples. We observe that, although correlation values are slightly different, the global patterns are preserved. These results, where correlation behavior remains stable over time, suggest to the specialist that pollution is not merely a current, isolated problem, but that it has been an ongoing issue.

The correlations were computed between the measurements obtained from the spectrum analyzers (the reference device) and the multi-sensor device (MOX), for each pollutant. We notice that values are not correlated, as confirmed by the correlation matrices shown in Fig. 11. ST_NO_2 and MOX_NO_2 have a correlation of 0.16; ST_NO_x and MOX_NO_x have a correlation of 0.66; and ST_CO and MOX_CO have a correlation of 0.89. However, there is a higher correlation between $ST\_C_6H_6$ and MOX_CO, which can be confirmed observing two visualizations: the line graphs in Fig. 12 and the time circular diagrams in Fig. 13.

In Fig. 12 we observe a high correlation between measurements of $ST\_C_6H_6$ and MOX_CO recorded over a short period (Fig. 12.a), and similar correlation patterns are observed over a longer 6-month period considering the aggregated data (Fig. 12.b). It is also possible to observe certain patterns in Fig. 12.a in the period May 18-21 for MOX_CO and in the period May 17-21 for $ST\_C_6H_6$. These patterns can be confirmed in the time circular diagrams computed for the same time period for variables MOX_CO (Fig. 13.a) and $ST\_C_6H_6$ (Fig. 13.b), which highlight days for which correlation measures are above a user-defined threshold, 0.8 in this case. For instance, we identify in both diagrams that correlation between hourly measurements each day is higher in May 18 to May 21.

The environmental engineer highlighted the resources provided by *TV-MV Analytics* that facilitate visual investigation of correlations in air quality measures considering both small and large time windows.

U.S. Companies Stock Market Data

The stock market dataset was extracted from Quandl[§], originally containing information about 3,000 companies. For this study we considered data on equities from 438 technology companies (as listed by the New York Stock Exchange[¶] and the Nasdaq Stock Market[‖]), from January 2, 1962 to November 30, 2016. From the seven original

[§]Wiki EOD Stock Prices, https://www.quandl.com/data/WIKI/documentation/bulk-download (December 28, 2017)

[¶]https://www.nyse.com/ (December 28, 2017)

[‖]http://www.nasdaq.com/ (January 13 2017)

variables, we selected to explore the variable *close* (the stock's closing price), recorded daily. This use case explores only one variable and therefore does not demand multivariate data analysis. However, it illustrates the scalability and flexibility of the *TV-MV Analytics* framework, as it supports handling extensive time series relative to over 400 different companies to compare their behavior regarding the attribute *close*.

Several authors proposed visualizations for stock market data analysis. Similar to the solution by Ziegler et al. (7), where segments of software/bank stocks are clustered into similar patterns, we investigated possible global patterns in the technology sector. Considering a segment of 438 technology companies, is it possible to identify groups of companies with globally similar behavior? We defined a query to retrieve the average monthly values of *close* from January 1, 1995 to November 30, 2016 (aggregated data values as described in Eq. 2) for the target companies. We created the corresponding time matrices for each company, and obtained their feature vectors using color moments as features. For clustering the feature vectors we employed the Euclidean distance as dissimilarity measure and the X-means clustering algorithm.

We explored the hierarchical small multiple time matrix views to identify representative behaviors over the observation period, considering the impact of missing data. Fig. 14.A shows the similarity map of variables, indicating the five clusters obtained with the X-means algorithm ($SC = 0.33$). The blue cluster (C1) includes 99 companies, the lilac cluster (C2) 60 companies, the red cluster (C3) 45 companies, the olive cluster (C4) 121 companies, and the green cluster (C5) 113 companies.

Considering these five clusters of companies with similar behavior over time, we take their corresponding medoids as companies representative of the characteristic temporal pattern of each cluster: TeleTech (TTEC), AOL, JDS Uniphase Corporation (JDSU), Rovi Corporation (ROVI), and Fortinet (FTNT); their corresponding time matrix views are shown in Fig. 14.B. This view corresponds to the first level of the hierarchical small multiples view. The user can browse to further inspect the time matrices relative to the companies within each cluster. As an illustration, we show in Figure 15 a possible view resulting from further inspecting the contents of each cluster. It shows the time matrix views of three arbitrarily selected companies from each cluster in Fig. 14. These examples illustrate that the groups are well-formed, including companies with similar temporal behavior of variable *close*.

In the blue cluster, low values of *close* occur mostly at the beginning of the observed period and higher values occur mostly towards the end. The behavior of the lilac cluster is characterized by a high percentage of missing data and extreme values, both high and low. The yellow group is characterized by reduced variability. The green cluster is characterized by a high ratio of missing data, and also presents extreme values. The red cluster exhibits a particular behavior where high values occur at the beginning of the period (initial rows) and low values predominate later; this behavior is observed for all companies in this cluster.

To illustrate the potential of the time circular diagram (TCD) to support the investigation of periodic correlation patterns we consider the behavior of IBM *close* over an extensive time period (from January, 1962 to November 2016). Each cell in the interactive TCDs represents a month in the original data (a)-(b) or the Haar-aggregated data (c)-(d), see Fig. 16. Lines relative to monthly values are joined into groups of 12 (user defined), defining one-year periods, where the lines are visually perceived as triangles. We explore the TCDs at four different levels of the Haar wavelet transformation: (a) and (b) show views of level 0 (i.e., the original data) and (c) and (d) show views of level 3, obtained aggregating the 12-month intervals. The correlations between different periods can be highlighted by selecting a target period, a functionality illustrated in the TCDs depicted in (b) and (d). In both, the 12-month period relative to year 2002 has been user-selected, and eight 12-month periods are identified as highly correlated (above 0.8) with year 2002 and highlighted in red. The years in which the monthly behavior of close have been found to be highly correlated with 2002 are 1962, 1966, 1968, 1970, 1973, 1981, 1991, and 2004. This is more clearly observed in Fig. 16(d), which uses a higher Haar compression level, than in Fig. 16(b), which shows the original data and lines are highly overlapped. Thus, the TCD views at higher Haar levels allow dealing with information overload, as strong correlation patterns are preserved and can be better perceived at higher compression levels.

Finally, the prediction functionality is illustrated with an experiment of predicting the behavior of variable *close* for a particular year, considering different intervals of known historical data (Eq. 9). Results are shown for the variable *close* of IBM, whose time matrix view is shown in Fig. 17. The more information the model has, the more accurate is the prediction (47). This is confirmed by the prediction results for 2016, obtained by considering distinct periods of historical data as input: 31, 15 and five years, shown in Fig. 17 in comparison with actual values. More precise results

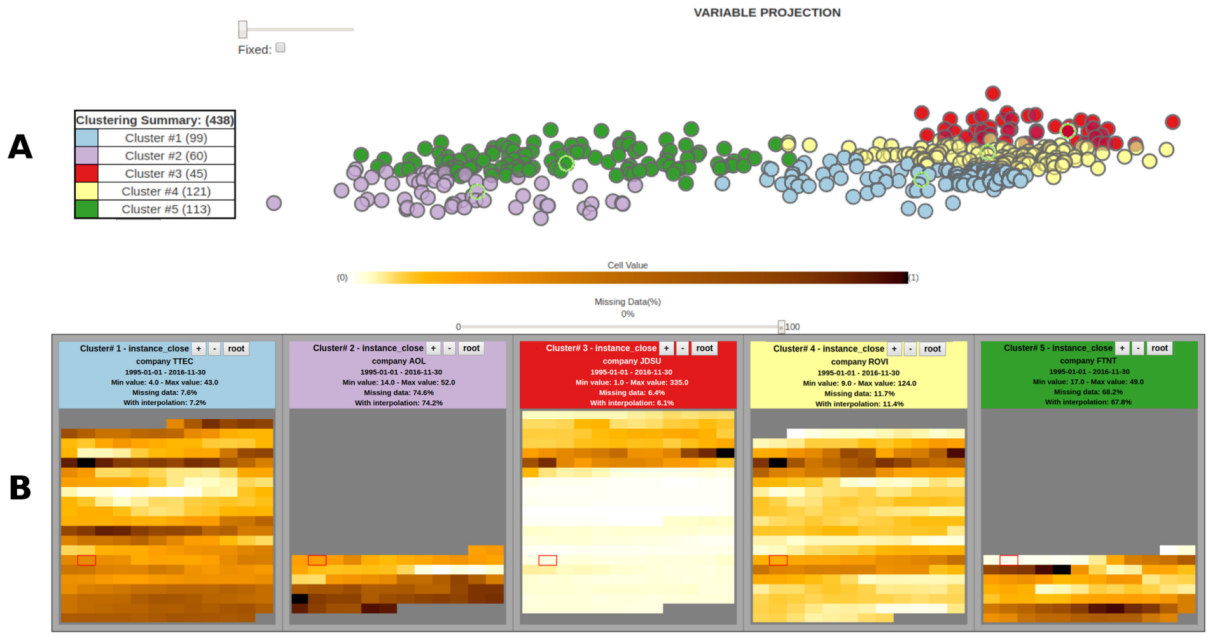


Figure 14. Hierarchical time matrices of the variable *close* for 438 companies (from January 1, 1995 to November, 30 2016): (A) IDMAP similarity map view, with cluster medoids highlighted in bright green; and (B) small multiple time matrix view of the variables that are the corresponding medoids of each cluster (TeleTech (TTEC), AOL, JDS Uniphase Corporation (JDSU), Rovi Corporation (ROVI), and Fortinet (FTNT)). Each matrix entry shows the maximum monthly value.

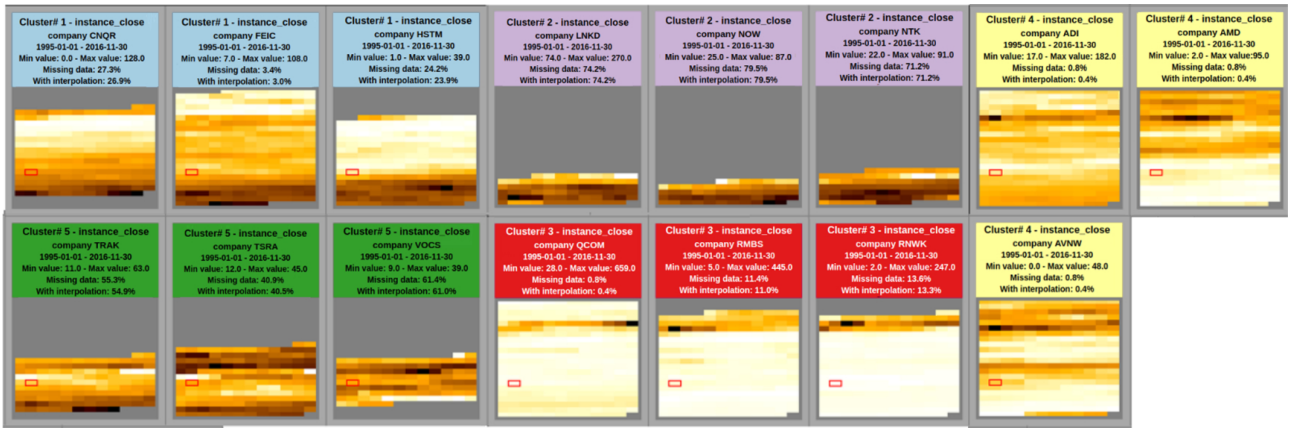


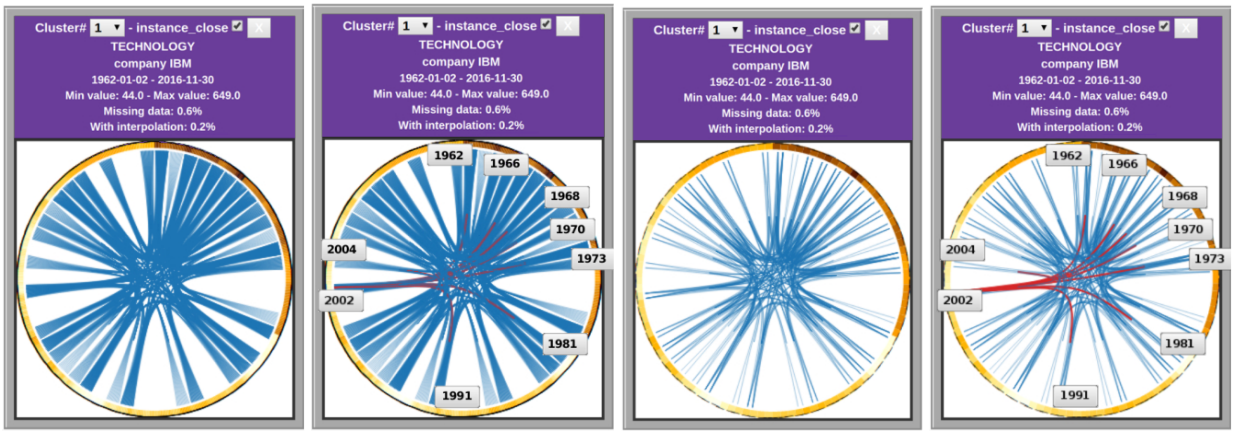
Figure 15. Time matrix views of variable *close* for three U.S technology companies randomly selected from each of the five representative clusters of companies identified in Figure 14 (TTEC, AOL, ROVI, FTNT, JDSU). They are illustrative of companies that can be accessed in the second level of the hierarchical time matrices visualization, upon user interaction. The examples confirm that the clusters illustrate groups with very different behaviors of *close*, but companies in a same cluster do show similar behavior.

were obtained considering a longer periods of historical data, as expected.

We discussed these scenarios with a business manager who is also a stock market investor. According to him, individual investors own approximately half of all stocks available on the U.S. stock market. Tools that can assist in the decision-making process are crucially important for stock selection. Many factors cause volatility in the stock market, and *TV-MV Analytics* supports the identification of companies that are stable or unstable over a long term, which is an important aspect for investors.

Conclusions and Future Work

We have introduced the visual analytics framework *TV-MV Analytics* supporting exploratory data analysis of complex time-varying multivariate data. The framework supports, in an integrative manner, feature extraction, clustering, individual and global visualizations of data variables and data instances. The supported analyses and visualizations convey a detailed representation of temporal behavior of data at multiple user-defined aggregation levels. The framework provides overviews of multiple variables regarding behavioral similarity for user-defined time periods and enables users to identify representative variables for characterizing a target phenomenon. We have



(a) TCD level 0

(b) TCD level 0 highlighting periods correlated with a selected period

(c) TCD in level 3

(d) TCD level 3 highlighting periods correlated with a target period

Figure 16. Time circular diagrams depicting the behavior of IBM *close* from 1962 to 2016, where each cell represents: a month in the original data (a)-(b), and aggregated data (c)-(d). The monthly values have been bundled in groups of 12 months (set by the user), where the lines are visually perceived as triangles. The TCD was explored at four Haar levels, where (a)-(b) depict the original data, and (c)-(d) shows the TCD at level 3 by aggregating the intervals in the user-defined groups of 12 months. Eight 12-month periods were identified as the most correlated to year 2002, namely 1962, 1966, 1968, 1970, 1973, 1981, 1991, and 2004.

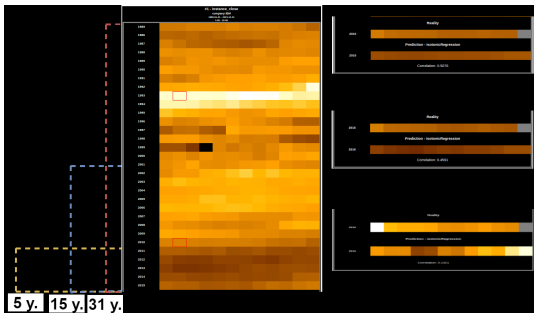


Figure 17. Prediction results of variable *close* for company IBM, obtained considering three alternative periods of the historical records available from January 1985 to December 2015.

conducted an experimental assessment of the capabilities and effectiveness of *TV-MV Analytics*, by considering time-varying multivariate data from three different domains: crime statistics indicators for the state of São Paulo, air quality measurements in an Italian city, and U.S. stock market prices.

The case studies are representative of the use of the system for gaining insight into the historical behavior of multiple variables. The framework can greatly assist an analyst to answer questions and uncover previously unknown relationships. *TV-MV Analytics* is useful when i) the number of variables is high; ii) there are defined target variables; and iii) there is interest in investigating behavior at different time scales and for different periods time.

The framework in its current form has limitations. For example, to obtain a solution it is necessary to derive informative feature vectors from time matrix representations, However, the task of selecting the appropriate moments for a given dataset is not straightforward, and an analyst is required to have knowledge about data properties. Establishing an appropriate dissimilarity measure for data is another complex and often domain-specific issue that cannot be addressd in a simple or automatic way.

A more extensive evaluation of the usability of the time circular diagram is desirable, as well as a systematic evaluation of the framework done by domain experts. Such evaluations are important to provide relevant information about the potential advantages and limitations of the framework and its supported techniques before its intuitive use is possible.

Acknowledgments

This work received financial support of the State of São Paulo Research Foundation (FAPESP) grants 12/24537-0, 15/12831-0 and 17/05838-3, and from the Brazilian National Research Council (CNPq) grant 301847/2017-7.

References

[1] Martin S and Quach T. Interactive visualization of multivariate time series data. *10th Int Conf on Foundations of Augmented Cognition: Neuroergonomics and Operational Neuroscience* 2016; 9744: 322–332.

[2] Zhang J, Ahlbrand B, Malik A et al. A visual analytics framework for microblog data analysis at multiple scales of aggregation. *Comp Graph Forum* 2016; 35(3): 441–450.

[3] Cook KA and Thomas JJ. *Illuminating the path: The research and development agenda for visual analytics*. National Visualization and Analytics Center, 2005.

[4] Engel D, Greff K, Garth C et al. Visual steering and verification of mass spectrometry data factorization in air quality research. *Trans on Vis and Comp Graph* 2012; 18(12): 2275–2284.

[5] Dasgupta A, Kosara R and Gosink L. Vimtex: A visualization interface for multivariate, time-varying, geological data exploration. *Comp Graph Forum* 2015; 34(3): 341–350.

[6] Machado V, Leite R, Moura F et al. Visual soccer match analysis using spatiotemporal positions of players. *Comp & Graph* 2017; 68: 84 – 95.

[7] Ziegler H, Jenny M, Gruse T et al. Visual market sector analysis for financial time series data. *Symp on Visual Analytics Science and Technology* 2010; 1: 83–90.

[8] Brehmer M, Lee B, Bach B et al. Timelines revisited: A design space and considerations for expressive storytelling. *IEEE Trans on Vis and Comp Graph* 2017; 23(9): 2151–2164.

[9] Soriano-Vargas A, Vani BC, Shimabukuro MH et al. Visual analytics of time-varying multivariate ionospheric scintillation data. *Comp & Graph* 2017; 68: 96–107.

[10] Livingston MA, Decker JW and Ai Z. Evaluating multivariate visualizations on time-varying data. *Vis and Data Analysis* 2013; 8654: 86541–865414.

[11] Aigner W, Miksch S, Schumann H et al. *Visualization of time-oriented data*. Springer Science & Business Media, 2011.

[12] Hochheiser H and Shneiderman B. Dynamic query tools for time series data sets: timebox widgets for interactive exploration. *Inf Vis* 2004; 3(1): 1–18.

[13] Keim DA, Schneidewind J and Sips M. Circleview: A new approach for visualizing time-related multidimensional data sets. *Working Conf on Advanced Visual Interfaces* 2004; 1(4): 179–182.

[14] Van Wijk JJ and Van Selow ER. Cluster and calendar based visualization of time series data. *Symp on Inf Vis* 1999; 1: 4–9.

[15] Lin J, Keogh E, Lonardi S et al. Visually mining and monitoring massive time series. *10th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining* 2004; 10: 460–469.

[16] Turkay C, Parulek J, Reuter N et al. Interactive visual analysis of temporal cluster structures. *Comp Graph Forum* 2011; 30(3): 711–720.

[17] Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Computational and Applied Mathematics* 1987; 20: 53 – 65.

[18] Liu X, Hu Y, North S et al. Correlatedmultiples: Spatially coherent small multiples with constrained multi-dimensional scaling. *Comp Graph Forum* 2018; 37(1): 7–18.

[19] Tufte ER. Envisioning information. *Optometry & Vision Science* 1991; 68(4): 322–324.

[20] Zhang D, Zhu L, Wang C et al. Timespiral, an enhanced interactive visual system for time series data. *2nd Int Conf on Inf Management* 2016; 2: 127–133.

[21] Yang J, Ward MO, Rundensteiner EA et al. Visual hierarchical dimension reduction for exploration of high dimensional datasets. *Symp on Data Vis* 2003; 03: 19–28.

[22] Yang J, Hubball D, Ward MO et al. Value and relation display: Interactive visual exploration of large data sets with hundreds of dimensions. *Trans on Vis and Comp Graph* 2007; 13(3): 494–507.

[23] Schreck T, Fellner D and Keim D. Towards automatic feature vector optimization for multimedia applications. *Symp on Applied computing* 2008; 23: 1197–1201.

[24] Ng CU and Martin GR. Automatic selection of attributes by importance in relevance feedback visualisation. *8th Int Conf on Inf Vis* 2004; 1: 588–595.

[25] Yuan X, Ren D, Wang Z et al. Dimension projection matrix/tree: Interactive subspace visual exploration and analysis of high dimensional data. *Trans on Vis and Comp Graph* 2013; 19(12): 2625–2633.

[26] Turkay C, Filzmoser P and Hauser H. Brushing dimensions; a dual visual analysis model for high dimensional data. *Trans on Vis and Comp Graph* 2011; 17(12): 2591–2599.

[27] May T, Bannach A, Davey J et al. Guiding feature subset selection with an interactive visualization. *Conf on Visual Analytics Science and Technology* 2011; 1: 111–120.

[28] Seo J and Shneiderman B. A rank-by-feature framework for unsupervised multidimensional data exploration using low dimensional projections. *Symp on Inf Vis* 2004; 1: 65–72.

[29] Seo J and Shneiderman2 B. Knowledge discovery in high-dimensional data: Case studies and a user survey for the rank-by-feature framework. *Trans on Vis and Comp Graph* 2006; 12(3): 311–322.

[30] Piringer H, Berger W and Hauser H. Quantifying and comparing features in high-dimensional datasets. *12th Int Conf on Inf Vis* 2008; 1: 240–245.

[31] Turkay C, Lundervold A, Lundervold AJ et al. Representative factor generation for the interactive visual analysis of high-dimensional data. *Trans on Vis and Comp Graph* 2012; 18(12): 2621–2630.

[32] Miksch S and Aigner W. A matter of time: Applying a data–users–tasks design triangle to visual analytics of time-oriented data. *Computers & Graphics* 2014; 38: 286–290.

[33] ping Tian D et al. A review on image feature extraction and representation techniques. *Int J of Multimedia and Ubiquitous*

Engineering 2013; 8(4): 385–396.

[34] Abo-Zaid A, Hinton OR and Horne E. About moment normalization and complex moment descriptors. *Pattern Recognition* 1988; 301: 399–409.

[35] Jain AK. Data clustering: 50 years beyond k-means. *Pattern recognition letters* 2010; 31(8): 651–666.

[36] Brewer CA, Hatchard GW and Harrower MA. Colorbrewer in print: a catalog of color schemes for maps. *Cartography and Geographic Information Science* 2003; 30(1): 5–32.

[37] Koh HC, Tan G et al. Data mining applications in healthcare. *J of healthcare information management* 2011; 19(2): 64–72.

[38] Levkowitz H. *Color theory and modeling for computer graphics, visualization, and multimedia applications*. Springer Science & Business Media, 1997.

[39] Pupyrev S, Nachmanson L and Kaufmann M. Improving layered graph layouts with edge bundling. *Graph Drawing* 2010; 6502: 329–340.

[40] Lepik Ü and Hein H. *Haar wavelets: with applications*. Springer Science & Business Media, 2014.

[41] Minghim R, Paulovich FV and Lopes A. Content-based text mapping using multi-dimensional projections for exploration of document collections. *Electronic Imaging* 2006; 6060: 60600S1–60600S12.

[42] Samet H. *The design and analysis of spatial data structures*, volume 199. Addison-Wesley Reading, MA, 1990.

[43] Paulovich F, Nonato L, Minghim R et al. Least square projection: A fast high-precision multidimensional projection technique and its application to document mapping. *Trans on Vis and Comp Graph* 2008; 14(3): 564–575.

[44] Arvate P and Souza AP. The fire-armed police effect: Evidences from a quasi-natural experiment in brazil. *FGV, Sao Paulo School of Economics* 2016; 429.

[45] De Vito S, Massera E, Piga M et al. On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario. *Sensors and Actuators B: Chemical* 2008; 129(2): 750–757.

[46] De Vito S, Piga M, Martinotto L et al. Co, no2 and nox urban pollution monitoring with on-field calibrated electronic nose by automatic bayesian regularization. *Sensors and Actuators B: Chemical* 2009; 143(1): 182–191.

[47] Noriega A, Blanco D, Alvarez B et al. Dimensional accuracy improvement of fdm square cross-section parts using artificial neural networks and an optimization algorithm. *Int J of Advanced Manufacturing Technology* 2013; 69(9-12): 2301–2313.