



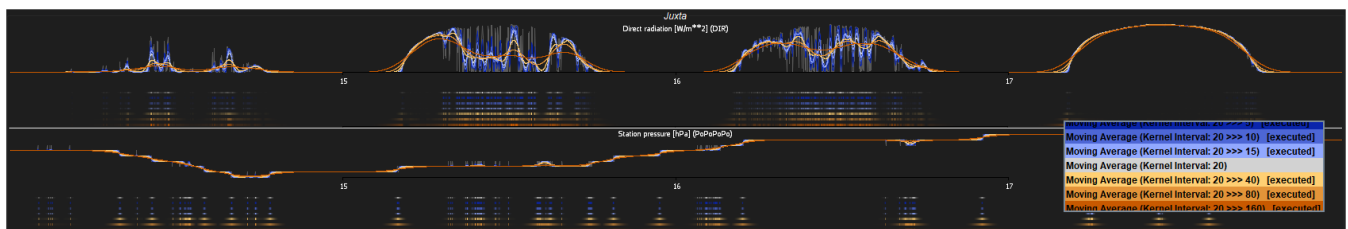
# Visual-Interactive Preprocessing of Multivariate Time Series Data

Jürgen Bernard<sup>1</sup> , Marco Hutter<sup>1</sup> , Heiko Reinemuth<sup>1</sup>, Hendrik Pfeifer<sup>1</sup>, Christian Bors<sup>2</sup>, and Jörn Kohlhammer<sup>3</sup>

<sup>1</sup>TU Darmstadt, Darmstadt, Germany

<sup>2</sup>TU Wien, Vienna, Austria

<sup>3</sup>Fraunhofer IGD, Darmstadt, Germany



**Figure 1:** Visual analysis of a moving average routine applied to multivariate time series (here: 2 dimensions). The uncertainty information over time introduced by the pre-processing routine is fed back into the approach (heat map) and can be related to the individual dimensions. Visual comparison of multiple parameterizations (encoded with color) allows the effective optimization of steering parameters.

## Abstract

Pre-processing is a prerequisite to conduct effective and efficient downstream data analysis. Pre-processing pipelines often require multiple routines to address data quality challenges and to bring the data into a usable form. For both the construction and the refinement of pre-processing pipelines, human-in-the-loop approaches are highly beneficial. This particularly applies to multivariate time series, a complex data type with multiple values developing over time. Due to the high specificity of this domain, it has not been subject to in-depth research in visual analytics. We present a visual-interactive approach for pre-processing multivariate time series data with the following aspects. Our approach supports analysts to carry out six core analysis tasks related to pre-processing of multivariate time series. To support these tasks, we identify requirements to baseline toolkits that may help practitioners in their choice. We characterize the space of visualization designs for uncertainty-aware pre-processing and justify our decisions. Two usage scenarios demonstrate applicability of our approach, design choices, and uncertainty visualizations for the six analysis tasks. This work is one step towards strengthening the visual analytics support for data pre-processing in general and for uncertainty-aware pre-processing of multivariate time series in particular.

## CCS Concepts

• **Mathematics of computing** → Time series analysis; • **Human-centered computing** → Visual analytics;

## 1. Introduction

Data pre-processing (PP) is the single most time-consuming phase in the entire data analysis pipeline. However, the characterization of this phase is enigmatic in most publications that address information visualization, visual analytics (VA), data mining, or machine learning. Many research projects use artificially clean data sources or focus on the parts of the pipeline that come after PP. At the same time, it seems obvious that the acceptance of data analysis technologies for the industry or practitioners is strongly influenced by the support of data PP, since this is where companies spend most of the time and cost. In essence, real-world data for solving real-world problems has to be pre-processed before decision making can start.

There are tools that specifically focus on data PP. Trifacta Wrangler is a service that supports the PP of typical business-related data sources that are fed to business analytics tools like Tableau, SAS, or Qlikview [BSS\*18]. Although time series require adequate processing approaches, these tools lack a considerable support for

time-oriented data, and even more so the PP of such. This paper further focuses on multivariate time series data (MVTS), which is even less supported with visualization and analysis techniques compared to univariate time series [AMST11]. Through their added complexity, MVTS exacerbate the users' problems related to data PP and analysis. Information visualization and VA can be excellent vehicles for data PP. To leverage the possibilities here, users need an effective pipeline to assemble and compare different PP routines using different parameters and yielding different uncertainty measures. But when using heterogeneous tools for PP, machine learning, and visualization, the VA process cannot adequately be supported. Our approach supports analysts and practitioners without extensive programming skills. This group of users benefit the most from interactive approaches and a direct visual reflection of changes to the PP parameters.

The rationale of this research effort is to systematically support the visual analysis of changes made to MVTS data by PP algo-

gorithms. We assume that every change to the data implies a potential drift-away from an original phenomenon encoded in the raw data. Ideally, analysts are easily able to identify the sweet spot between the degree of change made to the data and the preservation of relevant information. Just as well, many PP routines will explicitly accomplish the reduction of less important information in data to a more compact representation that still covers a targeted phenomenon in a meaningful way. We postulate that the uncertainty introduced with PP routines is a valuable source of information. In an optimal case, uncertainty information can be re-played into the analysis process to better support analysts with their tasks. In general, it would be desirable that PP pipelines for MVTS can be defined, validated, and refined in a human-centered, transparent, and trust-building way, similar to general VA principles [SSK\*16].

However, PP MVTS has not yet been subject to VA research. First, there is a gap between the rich set of routines for PP MVTS (applied in data mining, machine learning, and information retrieval communities [Fu11]) and interactive tools for the creation of pipelines for data PP. In particular, there is a lack of approaches that exploit such routines with VA principles. A second challenge is the dimensionality of the value domain, which makes MVTS particularly difficult to process, visualize, and analyze [AMST11]. With an increasing dimensionality, it becomes difficult for analysts to assess and compare the effects of routines on individual dimensions. The third challenge relates to the time-oriented uncertainty information that is introduced by the routines of PP pipelines. Even though many approaches leverage uncertainty information for informed decision making [WYM12], it is rarely considered a part of the data cleansing or wrangling process. Finally, analysts are confronted with the general challenge of model parameterization [SHB\*14]: Many routines have at least one parameter that needs to be tuned, but the comparison of multiple results processed with different parameterizations amplifies the challenge of comparing multiple MVTS visually.

We adopt VA principles for PP MVTS with assessment of uncertainty information in the process, which has not been subject to systematic research in the VA community so far. With this work, we want to bridge the gap between existing algorithmic PP routines and capabilities for the interactive construction of data processing pipelines. The four primary contributions are as follows:

- We discuss the design space for VA solutions for PP MVTS. For that purpose, we follow a task-based approach to address primary challenges associated with six core analysis tasks.
- We present design solutions for visual interfaces to address challenges associated with PP MVTS. These design solutions are based on the task characterization and result from discussions of alternative visual representation and layout design choices.
- We identify task-based requirements of toolkits for the interactive creation of PP pipelines for MVTS and briefly outline the applicability of existing baseline frameworks. The characterization supports designers and analysts in choosing or designing a toolkit, and in the interactive creation of PP pipelines, respectively. Finally, we present our approach that builds upon a toolkit that addresses all requirements.
- We demonstrate the applicability of the VA approach and justify the visualization designs in two usage scenarios that apply the toolkit to the visual-interactive PP of MVTS.

## 2. Related Work

After a review of concepts and techniques related to PP, we focus on approaches facilitating the visual-interactive analysis of MVTS. Finally, we briefly outline recent advances in uncertainty analysis.

### 2.1. Pre-Processing of Time Series Data

In general, PP data becomes necessary whenever the data does not match the requirements of downstream steps of analytical workflows. In the data mining, machine learning, and information retrieval, these downstream steps include content-based information retrieval [LSDJ06], indexing [Mül07], tracking [MHK06], similarity search [KTWZ10], feature analysis [Mör06], descriptor analysis [KK03], motif discovery [Fu11], anomaly detection [SMF15], rule discovery [Mör06], classification [MR06], clustering [WL05] segmentation [BBB\*18], labeling [BDV\*17], prediction [EA12], monitoring [LKL\*04], or exploratory search [Ber15]. In visualization approaches these downstream steps are additionally conflated with the goal to support users with effective visualizations [AMST11] requiring meaningful data preparation [SAAF18].

PP is considered the first fundamental step within the KDD reference workflow [FPS96] as well as for time series data mining and analysis [AA13, Ber15]. The special characteristics of time-oriented data [AMST11] require the individual treatment of the temporal and the value domain. Several surveys for time series PP exist, some of which elaborate taxonomies of PP algorithms [KK03, LKL\*04, Mör06, Fu11, Ber15]. We briefly review relevant classes of PP techniques, subdivided into cleansing and reduction, before we highlight visual-interactive PP approaches.

#### 2.1.1. Data Cleansing (Data Wrangling)

The determination of when data is clean is challenging since there is not one definition of clean data [KHP\*11]. One possible way to structure data cleansing is to differentiate between the *problem space* and *solution space*. Similarly, concrete techniques often consist of an error detection and handling component. In the following, we adhere to characterizations and taxonomies of dirty data that help analysts to structure the problem space [KHP\*11, GGAM12].

An important class of techniques is the detection and handling of *missing values* [SS18]. Imputing missing values is an approach that can involve linear and adaptive interpolation or regression techniques [BFG\*15]. Other error types include *implausible*, *ambiguous*, or simply *wrong* values. A related PP step is dealing with *outliers and anomalies*; in contrast to implausible values, outliers are considered plausible but may require special treatment. For the *reduction of noise*, solutions often employ moving average techniques with parameters for the kernel function [Mör06]. *Normalization* routines [KK03] help to make time series comparable and applicable for both analysis and visualization techniques. Cleansing the temporal domain often includes *equidistance* of time stamps, e.g., to foster data reduction approaches [BBGM17].

#### 2.1.2. Data Reduction

Most data reduction techniques eliminate irrelevant parts of the data while preserving relevant information. Compact data representations help to improve the performance and scalability of the subsequent analysis steps as well as visualization approaches [SAAF18]. Important classes of techniques include *sampling* and *filtering*, with the overall idea to reduce information with respect to a pre-defined criterion [Fu11]. As an alternative, *descriptors* can be applied to revive compact representations of the data [KK03, EA12].

### 2.1.3. Visual-Interactive Pre-Processing of Time Series

Several approaches advocate visual-interactive PP, building upon the human-centered challenge to assess data quality [LK06], the ability for visual inspection and direct manipulation [KHP\*11], or judgment in combination with automated computation [BAF\*13]. In this work, we apply these principles to PP MVTS. Inspiring was an approach for the visual-interactive PP of univariate time series [BRG\*12]. We borrow the idea to compare the input and output of PP routines as well as different outputs with varying parameters, as a basis for interactive navigation strategies and parameter space analysis [SHB\*14]. Beyond that, our approach includes a systematic characterization of design alternatives as well as uncertainty support. With the focus on visual-interactive techniques for the exploratory analysis of time series, the ChronoLenses approach is inspiring for this work as well [ZCPB11]. We adopt the juxtaposed and superimposed visualization of time series, and apply it to PP tasks. The Visplause [ASMP17] design study is one of the few VA approaches that also supports PP time series data, exhibiting inspiring visualization designs for MVTS and uncertainty assessment. Related to the latter is the “Know your enemy” tool [GE18], enabling users to identify and assess quality problems.

## 2.2. Visual Analysis of Multivariate Time Series

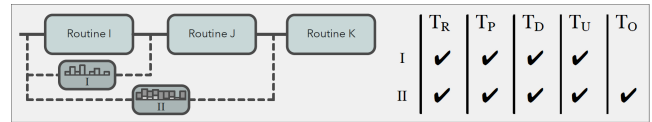
We review techniques for the visual analysis of MVTS. According to surveys and evaluations on visual comparison [EMJ10, GAW\*11], we divide relevant techniques into two categories.

*Small multiples* visualizations juxtapose individual dimensions to a list of charts with a shared x-axis (time). The value domains are either represented with position encodings (line charts, area charts, bar charts, symmetric area charts) [EMJ10, AMST11, CLKS19, LYK\*12, CLKS19] or encoded with color, opacity, shape, or stroke size (heat map approaches) [GTPB19, CLKS19], or both position and color in case of the horizon graph [Rei08]. The vertical display space often limits the number of dimensions that can be shown simultaneously [EMJ10]. The required y-space and four other factors build the basis for our characterization of the applicability of small multiples charts to encode PP uncertainty over time.

*Shared-space* or large singles visualizations build the second category. Line chart bundles and braided graphs [EMJ10, AMST11] use superposition to include multiple dimensions. Presumed that the span of the y-axis is small, this comparison technique is beneficial compared to small multiples [EMJ10]. With stacked line charts or stacked area charts [BW08, HHN00], we use a class of shared-space visualizations whenever the sum of individual value domains is of interest. Dimensionality reduction builds the basis for a shared-space variant with the principle to map high-dimensional data to a 2D path of time stamps [HWX\*10, BWS\*12, BSH\*16]. The technique scales for many dimensions and time stamps, and allows the visual exploration, comparison, and summarization of complex phenomena such as anomalies [BWS\*12], frequent patterns [BWK\*13], and periodicities [WVZ\*15]. Drawbacks of dimensionality reduction are the introduction of errors caused by the mapping, overplotting of similar time stamps, and axes which cannot be interpreted easily [SZS\*16].

## 2.3. Quantification and Visualization of Uncertainty

We ground our understanding of uncertainty in MVTS on probabilistic uncertainty modeling [BHJ\*14]. We employ a quantitative estimation of the uncertainty of each time stamp [WYM12, LS18]



**Figure 2:** Two requirements for PP MVTS: adding “visual sensors” everywhere in the pipeline, to conduct Single-Routine Assessment (I), or Multi-Routine Assessment (II). Both interactions enable all five analysis tasks, except T<sub>O</sub>, which requires (II).

from data processing routines. To support uncertainty-aware PP down to the finest granularity, our visualization designs require one quantitative uncertainty value for every timestamp  $t$ , dimension  $d$ , MVTS  $X$ , parameterization  $p$ , and PP routine  $r$ . To accomplish this, we apply a strategy for the quantification of uncertainty, which determines a normalized relative difference value for each timestamp and dimension  $u_{rel}(x_{t,d}) = \frac{abs(x_{t,v,r} - x_{t,v,r-1}) - \mu_z}{\sigma_z}$ , where  $\mu_z$  is the mean difference, and  $\sigma_z$  is the deviation [BBB\*19].

The incorporation, characterization, and evaluation of uncertainty in visualization methodologies and applications [BHJ\*14, Mac15, SSK\*16, BPHE17, HQC\*18] is a widely accepted subject to research. However, the analysis of uncertainty produced by processing algorithms along a pipeline is considered an open challenge [vLFR17]. Methodologies for multiple types of uncertainties for processing steps exist [WYM12], just like uncertainty visualization approaches for PP and quality assessment [BBB\*18, BBGM17, CCM09]. However, visual support for temporal uncertainty for PP is still rather uninvestigated. User evaluations and design studies on *temporal uncertainty* visualizations exist [GBFM16, WBFL17], but focus on changes in event states, rather than uncertainty time series over time. Our solutions depict uncertainty as an additional dimension for every dimension of the MVTS, similar to general visualization techniques for time series (cf. Section 2.2). For the aggregation of many uncertainty time series leading to (statistical) distributions over time, we employ bundles of line charts, quartile trend charts, as well as box plots over time [PWB\*09, AMST11, BHJ\*14, RBS\*18]. Finally, we derived valuable insights for characterization of visualizations from a user experiment on graphical perception of multiple time series [EMJ10], which we adopt for PP MVTS.

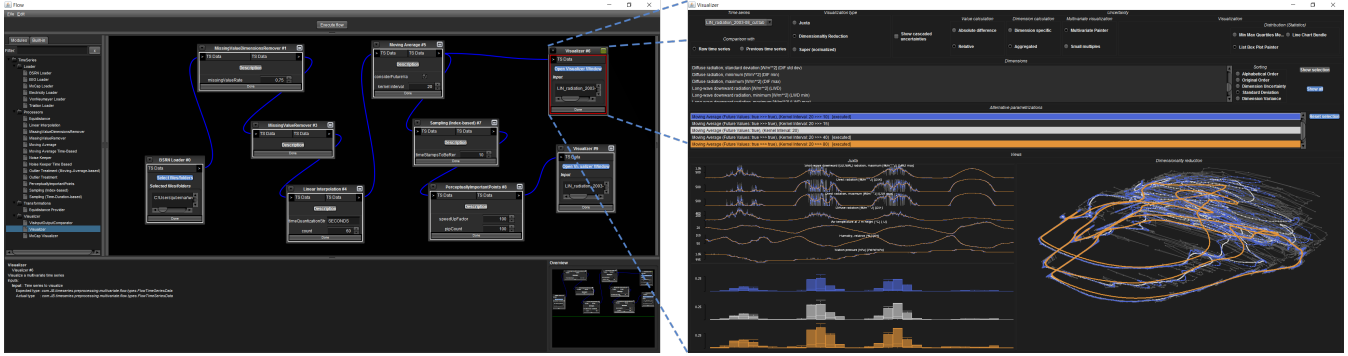
## 3. Approach

We present a VA approach for visual-interactive PP MVTS. Analysts, as well as practitioners without programming skills can interactively assemble and combine routines into a pipeline that can be customized and executed. We provide visual interfaces to support the analysis of the pipeline including intermediate results, and uncertainties at different levels of granularity. With the approach, analysts have means to exploit the rich set of existing PP routines [Fu11] to solve data quality problems and transform MVTS into useful forms for effective downstream analysis.

Our rationale was to follow a task-oriented design approach. Based on six primary analysis tasks (Section 3.1), we designed visual interfaces that support analysts to execute these tasks (Sections 3.2 - 3.7). To let readers benefit from the huge space of visualization designs, we characterize the visualization design space, evaluate alternative designs, and justify our design solutions.

### 3.1. Task Characterization

At a glance, we support analysts in the creation of pipelines, as well as in the interactive analysis of the output [vLFR17]. With the order



**Figure 3:** Visual-interactive tool for the creation of PP pipelines for MVTs (Usage Scenario 1). The left window shows the tool for the creation and modification of the pipeline ( $T_C$ ,  $T_O$ ): The tree view at the left contains a structured overview of the input- processing- and visualization modules. The center of the window shows the processing flow, in this case consisting of one module for loading, six PP routines as well as two “visual sensors” to interactively couple the pipeline with visualizations ( $T_R$ ). The visualization module in the window at the right window shows an intermediate result of the pipeline: The juxtaposed and the dimensionality-reduced visual representation for seven dimensions of the MVTs are shown side-by-side. Both techniques enable the visual comparison of input (dark gray) and output (light gray) MVTs ( $T_R$ ). In both cases, color facilitates the visual comparison of three different parameterizations (blue, light gray, orange) ( $T_P$ ). Juxtaposition is preferable for the detail-rich visualizations of few dimensions whereas dimensionality reduction helps to validate changes made to all dimensions at a glance. The uncertainty visualization at the bottom ( $T_U$ ) uses boxplot charts over time to scale for both: dimensions and time stamps.

of tasks, we adhere to the natural way of creating pipelines: from single routines to complex pipelines. The final overview task may then induce global-to-local navigation strategies [SHB\*14].

- $T_C$  **Creation:** Interactive construction and execution of a pipeline
- $T_R$  **Routine:** Assessment of effects of a single routine
- $T_P$  **Parameters:** Effects of alternative parameter values
- $T_D$  **Dimensions:** Output analysis for individual dimensions
- $T_U$  **Uncertainty:** Relate uncertainty of routine with MVTs
- $T_O$  **Overview:** Assessment of effects of routines in the pipeline

### 3.2. Interactive Creation of Pre-Processing Pipelines ( $T_C$ )

We define requirements of the baseline toolkit that should allow the construction of PP pipelines as well as the analysis of MVTs and discuss the applicability of a selection of well-known toolkits. Finally, we show a prototypical toolkit that meets all requirements.

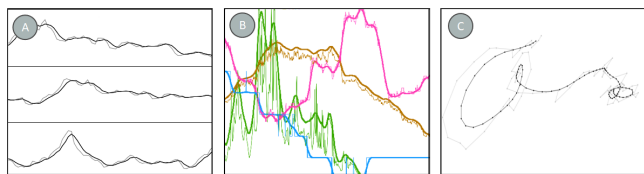
**Requirements met by existing Toolkits** We adopt important basic requirements from previous conceptual works and surveys [SHB\*14, vLFR17, BSS\*18]. Toolkits should allow analysts to combine different processing steps, build branches of data flows, and steer parameter values in order to explore the underlying data set and processing options. Our focus here is on MVTs, and there already exist libraries with dedicated data structures and PP routines for MVTs. The toolkit should therefore make routines from third-party libraries available and easily usable in the working environment. Due to the importance of visualization, it is crucial that this also refers to custom visualization components. Finally, we aim for a large user group being able to use the toolkit.

**Alternative Preprocessing Toolkits** These requirements are met, to various degrees, by several well-known toolkits for data mining, machine learning, and visualization. For example, WEKA [HFH\*09] offers rich sets of data mining and machine learning functionalities. KNIME [BCD\*07] is an analytics platform for data science and workflow creation. The Orange [DCE\*13] toolkit is a Python environment with a focus on data mining and visualization. RapidMiner [Rap] is an integrated machine learning workflow creation environment. In particular, a

common task that is already well supported by existing data analysis toolkits is the construction of a pipeline. The same is true for offline parameter steering and general visualization support.

**Requirements not yet met by existing Toolkits** There are task-oriented requirements for PP MVTs that go beyond what is currently offered by the well-known toolkits. These requirements, which have been generalized from related works (cf. Section 2.1.3 and previous projects in the areas of climate research and medical data analysis (cf. Section 4), are important for all of the characterized tasks ( $T_R, T_P, T_D, T_U$ , and  $T_O$ ). The task-oriented requirements endorse a closer coupling of the pipeline construction and visualization. This implies a form of *interaction* that goes beyond the pure construction and structural modification of the pipeline – namely, an interaction with the pipeline that supports the analysis of the effects of individual processing steps. To achieve this, the analyst should be able to interactively add “visual sensors” at any point within the pipeline, to achieve a whitebox nature of the entire workflow (see Figure 2). One requirement is to visually compare the input and output data of individual processing steps ( $T_R$ ,  $T_D$  – Single-Routine Assessment). Another requirement is to use visual sensors to assess the input and output MVTs of any arbitrary subsequence of the pipeline ( $T_O$  – Multi-Routine Assessment). Additionally, the coupling of the PP steps with visualizations should support analysts in gaining an understanding for the effects of different parameterizations ( $T_P$ ), as well as for the uncertainty of involved routines ( $T_U$ ). This is crucial to allow analysts to relate parameterizations and uncertainties with the corresponding routine, so that she can steer routines and interactively select the most suitable parameterization for the downstream processing pipeline.

**Our Approach** For the purpose of a conceptual presentation of the potential benefits of these interaction possibilities, we used a very simple open source library called javagl-flow [jav]. Although it does not offer a rich feature set like the toolkits mentioned above, it meets all requirements necessary for PP MVTs. It has a simple plugin concept for custom processing steps. It also supports a simple integration of our visualization design solutions for the visual comparison of MVTs and different parameterizations at different



**Figure 4:** Three classes of techniques for the visual input-output comparison of MVTS. A: small multiples for every juxtaposed dimension show input (gray) and output (black). B: superposition of multiple dimensions. The approach requires color-coding, thus, it is limited to  $7 \pm 2$  dimensions. Input-output comparison requires another visual encoding. C: dimensionality reduction of MVTS reveals a path metaphor in 2D, allowing the visual input-output comparison of all dimensions in a large-singles display.

points in the pipeline. Figure 3 shows one example PP workflow and a visualization that was created with the toolkit.

For the *coordination of views*, the interface offers the standard functionality of common workflow construction toolkits, namely to drag-and-drop processing modules into the workspace and connecting their outputs to the inputs of subsequent modules (TC). In addition, analysts can add uncertainty-aware visualization modules that may either show a current state of the MVTS within the pipeline, compare input and output MVTS of a routine, or compare input and output at different points in the pipeline.

### 3.3. Assessment of Effect of a Single Routine (TR)

Assessing the effect of algorithmic models is a general task in many VA environments [SSZ\*17]. Our line of approach is to superimpose the input and output MVTS to uncover these effects, allowing the user to judge if the result of a routine satisfactory. An additional design challenge is that of visualizing the effects on *multiple* dimensions simultaneously with only a limited amount of display space.

**Alternative Visualization Designs** The review of related works in Section 2.2 offers three classes of visualization approaches. Figure 4 indicates the design space with three prototypes, respectively. One strategy is *juxtaposition* of individual dimensions of a MVTS (left image). Visual comparison of input (gray) and output (black) is easily feasible as long as the number of dimensions is manageable with the available display space. *Superposition* of multiple dimensions in a large singles display is the second design strategy (center image). The number of distinguishable colors (or shapes) needed to encode individual dimensions is a limiting factor. A supplementary encoding (e.g., thickness or dashing) enables the comparison of input and output time series. Normalization of the value domains is necessary, with the price of losing the absolute value domains. *Dimensionality reduction* builds the basis for mapping all dimensions of the MVTS into 2D (third approach, right image). With only one path metaphor (gray, black) for all dimensions of the MVTS, the visualization technique is agnostic to the dimensionality. This visualization is well-suited for the exploration of phenomena represented in all dimensions of the MVTS (e.g., frequent patterns, periodicities, cf. Section 2.2), as well as for the observation of effects of operators that are applied to all dimensions of the MVTS simultaneously (e.g., noise reduction, smoothing). However, the visualization suffers from overplotting for many (similar) time stamps. In addition, input-output comparison in this abstract representation of MVTS lacks semantic interpretability and may require special expertise.

**Our Approach** We prefer the juxtaposed and the dimensionality-reduced visual representation of MVTS as they do not require color to discriminate individual dimensions (preserved for parameter comparison). As a default, we use juxtaposition as long as the number of (user-selected) dimensions is manageable with respect to the vertical display space [EMJ10]. As the dimensionality reduction approach is agnostic to the number of dimensions, it may be an interesting complement for multiple dimensions. In Figure 3 (right), a smoothing process is shown for seven dimensions. The juxtaposition still works well, but would run into scalability problems for considerably more dimensions. The dimensionality reduction technique works almost unaffected from dimensionality problems: analysts can easily infer a periodic pattern in the MVTS that has effectively been smoothed.

### 3.4. Effects of Alternative Parameter Values (TP)

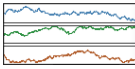
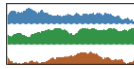
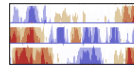
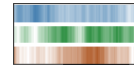
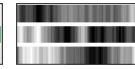
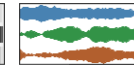
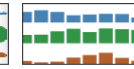
Many PP routines have at least one parameter that is steerable by the analyst [SHB\*14]. Without any guidance component, it is difficult to assess the effect of changes in parameter values, or even steer towards optimal parameter values.

**Our Approach** Our solution allows the visual comparison of multiple MVTS that are computed by a routine based on different parameterizations. The comparison of multiple results is hardly feasible with the class of superimposed visualizations shown in Figure 4, where categorical colors encode multiple dimensions in a joint value domain. Therefore, we enhanced our visual interfaces to support the comparison of multiple processing outcomes based on juxtaposition and dimensionality reduction. Figure 3 (right) shows the two visual-interactive interfaces. In both cases the use of color allows comparing the effects of different parameter values. We provide zooming and panning functionality for effective exploration, for the eventuality that the available space needs to be scaled.

### 3.5. Output Analysis for Individual Dimensions (TD)

One particular challenge related to MVTS is the dimensionality of the data. On the one hand, a visual interface is needed that allows the detailed analysis of individual dimensions on demand. On the other hand, analysts need to face the challenge to determine which of the available dimensions require manual inspection. In many cases, it is not effective to assess effects of routines for each available dimension in detail. However, there is always the risk of overlooking dimensions that may require manual inspection. Analysts may have different intents to conduct analysis of individual dimensions. It can be important to validate whether a routine produces useful results down to singular dimensions, or that a routine affects the individual dimensions appropriately. In some cases the characteristics of MVTS dimensions can be similar, like the ECG measurements used in Usage Scenario 2. Especially for heterogeneous dimensions such as the weather sensors used in Usage Scenario 3, assessing the range of effects may be crucial. Finally, depending on the dimension characteristics, routines can be applied globally, or might be parametrized differently.

**Our Approach** We support analysts with a component that directs the analysis towards interesting dimensions in the MVTS. Different interestingness scores are calculated by a degree-of-interest function. While our set of interestingness measures is not exhaustive, based on our insights into the application domain, we provide measures to evaluate dimensions by a) uncertainty, b) degree of change

							
	Line Chart	Area Chart	Horizon Gr.	Heat Map	Gray Scale	Symm. Area	Bar Chart
Required Y-Space	-	-	∅	+	+	∅	-
Temporal Resolution	+	+	+	+	+	+	-
Need for color	+	+	-	+	+	+	+
Line Chart Conflict	-	∅	+	+	+	+	+
Dimension Localization	+	+	-	+	-	+	+

**Table 1:** Applicability of small multiples chart types to encode temporal uncertainty information for multiple dimensions. The ordinal scores range from bad to good ( -, ∅, + ) used to characterize the applicability for different design aspects.

to the value domains, c) variance of the time series, d) original order, and e) the alphabetical order (to ease analyses addressing the semantic meanings). The scores of the ranked degree-of-interest functions are used as directive guidance [CGM\*17] to MVTS dimensions and allow the effective dimension selection. In the juxtaposed visualization of dimensions in Figure 4, the variance criterion was used for the dimension selection (high variances at the top). It can be seen that the usefulness of the smoothing routine differs between dimensions with high and low variance.

### 3.6. Relate Uncertainty of Routine with MVTS ( $T_U$ )

We postulate that uncertainty stemming from PP is a valuable source of information that can be replayed back into the analysis process. Our uncertainty quantification strategy measures the uncertainties introduced by PP routines [BHJ\*14], by calculating the normalized relative difference uncertainty for every timestamp, dimension ( $T_D$ ), parameterization ( $T_P$ ), and routine ( $T_R$ ) [BBB\*19]. To enable analysts to assess the uncertainty for multiple dimensions or parameterizations, we aggregate uncertainty into statistical distributions. Finally, summing up uncertainties of cascading routines supports the uncertainty-aware overview of the pipeline ( $T_O$ ).

**Alternative Visualization Designs** From a methodological perspective, we support relation-seeking between the MVTS and the uncertainty using the temporal domain as a shared primary key. Analysts will be able to identify uncertainty effects and link them to the phenomena in the MVTS just like the other way around. The design challenge at hand is the combined visualization of MVTS and uncertainty: For every dimension of a MVTS, we provide an


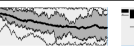

accompanying uncertainty time series. To account for design alternatives, we provide an overview of uncertainty visualizations for single uncertainty time series in the sense of small multiples in Table 1, i.e., *Line Chart*, *Area Chart*, *Horizon Graph*, *Heat Map*, *Gray Scale*, *Symmetric Area Chart*, and *Bar Chart* (cf. Section 2.2). For the overview of statistical distributions of uncertainty over time (large single techniques) in Table 2, we employ *Line Chart Bundles*, *Quartile Trend Charts*, as well as *Box Plot Charts* (cf. Sections 2.2 and 2.3). Inspired by the empirical works of Elmqvist et al. [EMJ10], Gschwandtner et al. [GBFM16], and Wunderlich et al. [WBFL17], we characterize the applicability of the design alternatives with respect to PP tasks, as well as visual scalability, temporal resolution, and perceptual (within-the-bar-bias [NS12]) issues, which designers may deem important in the connection of uncertainty-aware PP MVTS.

**Our Approach** We suggest the use of *Symmetric Area Charts* as a default for small multiples techniques (cf. Figures 8 and 9). If the span of the y-axis is small, *Heat Map* approaches are an effective alternative, which in theory only need one pixel of height (cf. Figure 1). Both techniques do not show any weaknesses in our characterization (see Table 1). For visualizing uncertainty distributions, we prefer the *Quartile Trend Chart* (Figure 7) if the number of time visualized stamps does not cause visual scalability problems. To address visual scalability issues, our default is the *Boxplot Chart* aggregating both the value and the temporal domain. Both techniques outperform the *Line Chart Bundle* technique (see Table 2).

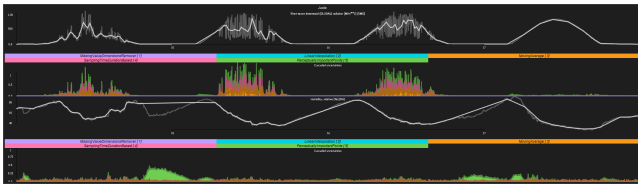
### 3.7. Assessment of Effects of Routines in the Pipeline ( $T_O$ )

With the baseline toolkit and the visual interfaces, analysts have the means to add, analyze, optimize, and connect individual routines. The visual-interactive nature of the toolkit also allows gaining an overview of resulting pipelines as a whole. Complementary to the routine overview is the overview of the impact of routines when the pipeline is applied to MVTS. Analysts may want to assess and compare the *effects* that individual routines of a pipeline have on the MVTS. It may be particularly beneficial to identify routines that cause the largest changes in the MVTS, or introduce the largest uncertainty. With these insights gained from the global overview, the baseline toolkit can then be used to navigate to routines that require improvement.

**Our Approach** We use the introduced uncertainty of every routine ( $T_U$ ) in the pipeline. Figure 5 shows the visual interface that provides an overview of effects of all routines ( $T_O$ ). We employ stacked area charts for the visualization of uncertainty of routines over time. This allows the assessment of both the overall amount of uncertainty as well as the individual uncertainties of routines.

			
	Line Chart Bundle	Quartile Trend Cht.	Boxplot Chart
Visual Clutter	-	∅	+
Required Y-Space	-	+	+
Aggregation Loss	+	-	-
Temp. Resolution	+	+	-
Temp. Scalability	-	-	+
Line Chart Conflict	-	∅	+
Within-the-Bar-Bias	∅	∅	-

**Table 2:** Applicability of shared-space chart types to encode a set of uncertainty series over time. The ordinal scores range from bad to good ( -, ∅, + ) used to characterize the applicability for different design aspects.



**Figure 5:** Overview of temporal uncertainties introduced by five routines of a pipeline ( $T_O$ ). Stacking of the uncertainties of the five color-coded routines allows the uncertainty-aware analysis of individual routines as well as the overall pipeline ( $T_U$ ). In the example, the most uncertainty is introduced with the moving average (orange), the sampling (pink), and the perceptually important points (green). Large green areas of uncertainty in the lower dimension indicate that there is room for improvement of the respective parameterization ( $T_P$ ).

Specifically, it can be determined if particular temporal regions were affected more significantly than others and if this can be attributed to a single routine. The green routine in Figure 5 added a considerable amount of uncertainty for the lower dimension. It may be advisable to investigate the routine in detail and improve the steering parameters accordingly ( $T_P$ ).

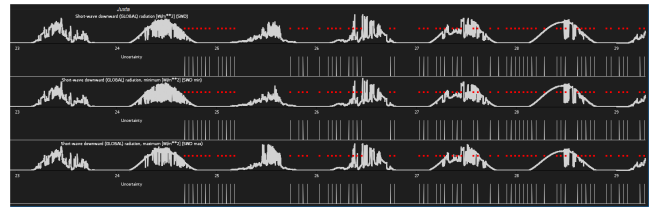
#### 4. Usage Scenarios

We demonstrate the applicability of our approach in two usage scenarios with different analysis goals. For the two scenarios, we use photovoltaics data with solar-dependent sensors [BKS12] and a climate data set measured in Antarctica [RLKL12].

##### 4.1. Reduction of Photovoltaics Data for a Web System

Compact and faithful reductions of MVTS data are a prerequisite for many algorithms and visualizations. In this usage scenario, we demonstrate how a complex MVTS is pre-processed into a compact representation that can be exchanged in a web-based client-server architecture. The data at hand is from a photovoltaics database with 20 solar-dependent sensors including different types of radiation measurements [BKS12]. Overall, 57 stations in the network measure these parameters all over the world allowing the assessment where on Earth photovoltaics technology could be most effective. With the temporal resolution (quantization) of 60 seconds every station produces 1,000,000 values every month. For being useful in a web-based data visualization context an effective data reduction strategy is necessary. This usage scenario reproduces a real-world use case that was conducted 2013 [BDF\*15], without having VA support. Overall, the creation, validation, and optimization of the PP had taken about two years. The PP process was scattered across multiple tools and systems, making parameter changes hard to track, and data had to be re-loaded constantly. With the tool, this is now resolved. Re-creating and validating the entire PP pipeline required less than one hour.

Together with the analyst, we start an early visual analysis of all dimensions. For that purpose, we add the data loader to the pipeline ( $T_C$ ), followed by a component for the visualization of the raw MVTS. One dimension only contains missing values, possibly because of a broken sensor. We remove this dimension and focus on missing values of the 19 remaining dimensions ( $T_D$ ). The red dots in Figure 6 disclose that the MVTS contains occasional missing values. We apply a linear interpolation routine to the pipeline ( $T_C$ ) to impute the missing values. An optional step may be to guarantee



**Figure 6:** Missing value replacement with an interpolation routine ( $T_R$ ). Zooming to an interval of 6 days for three dimensions reveals a series of missing values (red dots). The uncertainty visualization shows where missing values were imputed ( $T_U$ ).

the equal quantization after the changes made to the temporal domain. In such a case an equidistance routine with the natural quantization of 60 seconds may be appropriate.

Figure 6 also shows that the MVTS has a natural periodicity. The dimensions ( $T_D$ ) show regular increases and decreases as in the interval of six days (daily patterns). However, the dimensions contain a considerable amount of noise accompanying the periodic phenomenon. The analyst explains that most of the noise can be explained with varying cloud conditions which have an influence on solar radiation. However, for the web-based photovoltaics scenario it is more relevant to communicate the total amount of radiation rather than the occurrence of noise. Together with the analyst, we add a smoothing routine ( $T_{C,R}$ ) to make the data more useful. By using the variance-based dimension selection for guidance (cf. Section 3.5), we select 7 of the 19 dimensions with considerably differing variances (cf. Figure 3). With the input-output comparison technique, we zoom to a time interval that reveals the strong effect of the smoothing routine to the more variant dimensions at the top ( $T_R$ ). The dimensionality reduction approach at the right reveals the periodic phenomena in the data, which should be well-preserved after smoothing. To check the introduced uncertainty, we add an uncertainty view, where the analyst identifies significant spikes, corresponding to the regions affected by smoothing. By direct comparison, the orange parameterization introduces the most uncertainty ( $T_U$ ).

Informed by the visualizations, the analyst aims for parameter analysis to optimize the effect of the routine to the MVTS for the web scenario ( $T_P$ ). The analyst selects two dimensions for a fine-grained analysis of parameterizations (Direct Radiation and Station Pressure) shown in Figure 1. Overall, results with seven different parameter values can be compared (from 5 to 160 minutes, blue to orange). Just as well, uncertainty information introduced by different parameterizations can be assessed with the compact heat map technique underneath (cf. Section 3.6). The analyst infers that the averaging effect with orange parameter values is too coarse ( $T_P$ ) and decides for a moving average kernel of 20 minutes. This poses a good trade-off between signal preservation and noise reduction.

The smoothed MVTS is now in a state to apply data reduction ( $T_C$ ). The challenge at hand is adding a sampling routine ( $T_C$ ) to remove a considerable number of time-value pairs without losing too much information. With the dimension-selection support ( $T_D$ ), we select six dimensions with high variations of uncertainty (very high on top) and apply the sampling routine. In Figure 7 the dimensions are shown via juxtaposition. At the bottom the Quartile Trend Chart shows the aggregated uncertainty across all dimensions ( $T_U$ ). The analyst decides for a temporal kernel of 10 minutes: the routine then preserves the original phenomenon, while the temporal resolution yields a considerably more compact representation ( $T_P$ ).



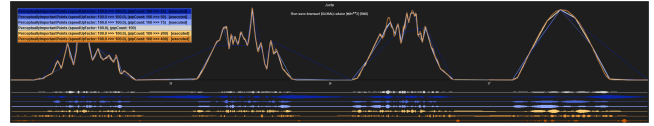
**Figure 7:** Assessment of the effect of a sampling routine to six selected dimensions ( $T_R$ ,  $T_P$ ,  $T_D$ ). The Quartile Trend Chart at the bottom depicts the uncertainty for every parameterization (aggregated over dimensions, cf. Table 2). The blue parameterization (5 minute kernel) almost introduces no uncertainty, in contrast to considerable changes caused by the orange kernel (40 minutes) ( $T_U$ ).

With the goal to further compress the MVTs, the analyst adds another data reduction technique: the perceptual important points (PIP) algorithm [ZJK10] preserves time-value pairs that are standing out and removes non-descriptive time-value pairs. As such, this technique is well-suited to prepare the data for the web-based photovoltaics use case. Figure 8 demonstrates how our small-multiples visualization supports the parameterization of the PIP routine for a single dimension ( $T_R$ ,  $T_P$ ,  $T_D$ ). Zooming to a time interval of four days reveals that 5,750 original time stamps could be reduced to a representation with only 100 time stamps (98% compression). With the result, we have achieved a representation of the MVTs that can be used for the web-based analysis system for the analysis of photovoltaics data.

The final PP pipeline is shown in Figure 3 (left). A data loader, six PP routines, and several “visual sensors” have interactively been added and combined. As a final task, the analyst wants to gain an overview of the effects of the entire pipeline to validate the result ( $T_O$ ). Figure 5 shows the input (gray) and output (white) time series for two selected dimensions ( $T_D$ ), as well as uncertainties from all employed routines. In addition to the visual comparison of the input and output MVTs, the analyst uses the uncertainty visualization to assess the effects of individual routines, as well as of the entire pipeline. The analyst observes that the PIP routine (green) introduced a significant amount of uncertainty for dimension at the bottom. This is not acceptable and must be addressed by re-configuring the PIP routine with a different parameterization ( $T_{R,P}$ ), followed by a step for the repetitive validation of the pipeline ( $T_{O,U}$ ).

#### 4.2. 30 Years of Climate Observation Data

The goal of this usage scenario is to make a longitudinal climate data set usable for effective downstream analysis. Since March 1981, a meteorological observatory program has been carried out at Neumayer Station (NM) ( $70^{\circ}37'S$ ,  $8^{\circ}22'W$ ), located in Antarctica. NM is an integral part of many international networks, e.g., the World Meteorological Organization (WMO). We use a data set that was recorded at NM over 30 years [RLKL12, BKS14]. The data was recorded every three hours and includes four sensors (air pressure, air temperature, wind direction and wind speed). In contrast to weather analysis, which usually focuses on time intervals of several days or weeks, the analysis of climate data usually is at the granularity of years. To support effective climate analysis, two challenges have to be faced with PP. First, the data needs to be cleaned to guarantee that data is usable by downstream data analysis algorithms. Second, data reduction is needed that will allow effective analysis but preserves aspects relevant for climate analysis.



**Figure 8:** Calibration of the perceptual important points (PIP) algorithm to achieve a compact and still representative data reduction ( $T_R$ ,  $T_P$ ,  $T_U$ ). A single dimension was selected to demonstrate the parameter calibration. For the time interval of four days (5760 time-value pairs in the original data), the PIP algorithm requires about 100 time-value pairs to represent the signal in a perceptually similar way (overall compression of 98%).

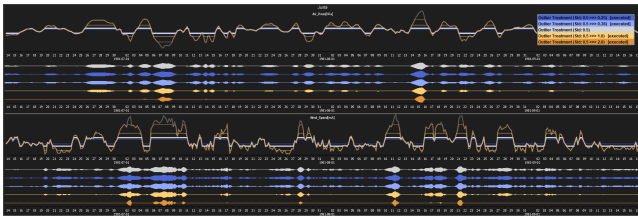
The analyst uses the small-multiples visualization to gain an overview of the raw weather data set with four dimensions (air pressure, air temperature, wind direction and wind speed). All dimensions contain fine-grained variation that can be omitted for long-term climate analysis. As expected, the analyst identifies a yearly pattern when investigating the air temperature dimension ( $T_D$ ). However, the MVTs contains missing values distributed across the temporal domain, which impede downstream data processing.

Hence, the analyst begins PP with the removal of missing the values and adds two routines to the pipeline ( $T_C$ ): The first routine simply removes missing values, whereas a second linear interpolation routine fills the gaps which are scattered across the value domain. Another step to achieve a certain level of data quality is to guarantee an equal quantization. In this connection, the analyst also aims at providing equidistance which is a pre-condition for many data mining and machine learning algorithms (cf. Section 2.1.3). For the parameterization of the equidistance routine, there is no need to fall back to our visual interfaces, as knowledge of the analyst about her domain and the data set requires a fixed quantization of 3 hours. In fact, using the uncertainty-aware small multiples visualization helps to validate that the routine did not cause any unexpected and unwanted violations to the usefulness of the MVTs.

The next step particularly requires visual interfaces, not only for validation but also parameterization purposes: The analyst wants to treat outliers to further improve data quality. The analyst experiments with a standard deviation-based routine, which crops the value domain whenever the distance of values to the mean exceeds a certain threshold. However, due to the natural periodicity of the data, the routine always crops day and night peaks: the interface in Figure 9 clearly demonstrates that the routine would cause harm to the MVTs ( $T_{R,D,P}$ ). It becomes apparent that the statistical outlier treatment algorithm does not take the temporal progression into account appropriately, which leads to sub-optimal results.

For climate analysis, the analyst wants to condense the fine-grained temporal information to a more robust representation. With the natural periodicity of one day, sampling the MVTs with a daily granularity appears to be a sensible choice ( $T_{R,P}$ ). To achieve representative sampling points the remaining challenge is smoothing the MVTs as an intermediate step. The analyst tests a moving average routine with parameterizations from three hours to eight days and starts to fine-tune the parameterization. This process is similar to the decision-making process in the first scenario, depicted in Figure 1. In this case, the parameterization of two days seems to be the sweet spot between the coverage of short-term weather phenomena and smoothing the data towards long-term climate analysis. After smoothing, the MVTs is now applicable for data sampling to receive a more compact data representation. The analyst adds





**Figure 9:** Statistical outlier treatment routine working in terms of distances of standard deviations to the mean. The routine cuts low and high peaks, regardless of the time-orientation of the data, which may be very error-prone for time-oriented data.

a sampling routine to the pipeline ( $T_C$ ) and starts to fine-tune the temporal kernel interval. After the analysis of alternative parameterizations and uncertainty effects, she decides to condense the MVTS to one value per day.

As a whole, the original data set with about 90,000 time stamps was condensed to a MVTS with about 11,000 time stamps (about 88% compression) and can now be used to conduct effective climate analysis. As a final step, the analyst wants to assess the potential fail that would have been caused by the neglected outlier treatment routine described earlier and described in Figure 9. For that purpose, the analyst uses two “visual sensors” and attaches the output of the neglected outlier routine with the final output (after sampling). At the right of Figure 10, it can be seen that the analyst was well-advised to neglect the outlier routine: It can be seen that the outlier treatment routine (gray) would have cut considerable parts of the valuable peak information. At the left of Figure 10, the final pipeline can be seen, consisting of a data loader, four PP routines, as well as several “visual sensors” that have successfully been used for the validation of routines, and routine parameters.

## 5. Discussion and Future Work

We identified possible future research directions that aim at providing a fully integrated PP and analysis suite for MVTS.

**Parameterizations for Individual Dimensions** Our interfaces allow visualizing individual effects that PP steps have on MVTS. We specifically address the problems of visualizing the effects on the individual *dimensions* of the MVTS, and on results of different parameterizations of routines. For the analyst, this may bring some important insights; namely, that the effect of a processing step with a certain parameterization is as desired for one dimension, but should be applied with a different parameterization for another dimension. The potential benefits and possible integration options for steering the parameterization of PP steps for individual dimensions still has to be investigated.

**Facilitating Data Integrity** Our contributions are one step towards an analysis suite for MVTS that makes PP an integral part of the VA workflow. But although PP often aims at ensuring certain consistency and quality criteria for MVTS, they may also have the opposite effect. For example, a downstream routine may require (or worse: silently assume) equidistant MVTS. If some upstream step only removes a single entry in a MVTS, this may cause this precondition to be violated. Properly stating and formalizing the actual requirements that certain analysis steps have for MVTS is difficult, but is an interesting area of research for the information visualization community in general. Therefore, describing, identifying, and resolving inconsistencies as well as establishing or maintaining certain *integrity criteria* for MVTS is subject to future work.



**Figure 10:** Final pipeline applied on the longitudinal climate data set. The analyst has put two “visual sensors” to the output of the OutlierTreatment and Sampling routine to facilitate the visual comparison of different branches. Right: the comparison reveals that the outlier routine (gray) would have cut considerable parts of the peaks, which have been preserved with the final branch (white).

**User Evaluations of Design Alternatives** We have characterized the extent of the design space for uncertainty-aware PP of MVTS and argued about particular design choices made when we demonstrated our visualization techniques. This characterization may also foster future user evaluation approaches to identify interesting experiment factors. In the scope of evaluation endeavors, we have a concrete agenda to conduct a field study with real-world users to test the effectiveness and efficiency of alternative uncertainty visualizations for different tasks. Also interesting is the assessment of trade-offs between design alternatives with regard to visual scalability to foster adaptivity (cf. Sections 3.3 and 3.6).

**Dimensionality Reduction** From the related work, we borrowed the idea to map phenomena of MVTS into paths in 2D as an interesting method for thinking about scalability. In addition, we confirm benefits for anomaly detection, noise reduction, and smoothing routines (see. Figure 3, right). However, dimensionality reduction suffers from a lack of interpretability, e.g., regarding the meaning of display axes [SZS\*16]. It would be interesting to evaluate how users interpret the dimensionally reduced plots.

## 6. Conclusion

We presented a novel VA approach that allows for the visual-interactive and uncertainty-aware PP of MVTS for the first time. For six analysis tasks, we discussed visualization design alternatives, presented uncertainty-aware visualization techniques and coupled these solutions with a visual-interactive prototypical toolkit that meets all our requirements to PP MVTS in the sense of a whitebox integration. With the approach, we aimed to bridge the gap between existing PP routines and general toolkits for pipeline creation. In two usage scenarios, we demonstrated that complex PP pipelines for MVTS can now be created with VA principles by a broader user group beyond data scientists. Future approaches may benefit from the characterizations of design considerations. This regards the selection or design of PP tools, visualization techniques for MVTS, as well as uncertainty-aware analysis interfaces. Finally, we have extended the space for user studies in connection with the proposed analysis tasks and visualization designs for PP MVTS.

## Acknowledgments

This work was supported by the Deutsche Forschungsgemeinschaft (DFG) and the Austrian Science Fund (FWF), Project No. I 2850 (-N31), Lead Agency Procedure (D-A-CH) “Visual Segmentation and Labeling of Multivariate Time Series (VISSECT)”.

## References

- [AA13] ANDRIENKO N., ANDRIENKO G.: A visual analytics framework for spatio-temporal analysis and modelling. *Springer Data Mining and Knowledge Discovery* 27, 1 (2013), 55–83. doi:10.1007/s10618-012-0285-7. 2
- [AMST11] AIGNER W., MIKSCH S., SCHUMANN H., TOMINSKI C.: *Visualization of Time-Oriented Data*, 1st ed. Human-Computer Interaction. Springer Verlag, 2011. doi:10.1007/978-0-85729-079-3. 1, 2, 3
- [ASMP17] ARBESSER C., SPECHTENHAUSER F., MÜHLBACHER T., PIRINGER H.: Visplause: Visual data quality assessment of many time series using plausibility checks. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 23, 1 (2017), 641–650. doi:10.1109/TVCG.2016.2598592. 3
- [BAF\*13] BÖGL M., AIGNER W., FILZMOSER P., LAMMARSCH T., MIKSCH S., RIND A.: Visual analytics for model selection in time series analysis. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 19, 12 (2013), 2237–2246. doi:10.1109/TVCG.2013.222. 3
- [BBB\*18] BERNARD J., BORS C., BÖGL M., EICHNER C., GSCHWANDTNER T., MIKSCH S., SCHUMANN H., KOHLHAMMER J.: Combining the Automated Segmentation and Visual Analysis of Multivariate Time Series. In *EuroVis Workshop on Visual Analytics (EuroVA)* (2018), Eurographics. doi:10.2312/eurova.20181112. 2, 3
- [BBB\*19] BORS C., BERNARD J., BÖGL M., GSCHWANDTNER THERESIA KOHLHAMMER J., MIKSCH S.: Quantifying Uncertainty in Multivariate Time Series Pre-Processing. In *Review. EuroVis Workshop on Visual Analytics (EuroVA)* (2019), Eurographics. submitted. 3, 6
- [BBGM17] BORS C., BÖGL M., GSCHWANDTNER T., MIKSCH S.: Visual Support for Rastering of Unequally Spaced Time Series. In *Visual Information Communication and Interaction (VINCI)* (2017), ACM, pp. 53–57. doi:10.1145/3105971.3105984. 2, 3
- [BCD\*07] BERTHOLD M. R., CEBRON N., DILL F., GABRIEL T. R., KÖTTER T., MEINL T., OHL P., SIEB C., THIEL K., WISWEDEL B.: KNIME: The Konstanz Information Miner. In *Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)* (2007), Springer. 4
- [BDF\*15] BERNARD J., DABERKOW D., FELLNER D. W., FISCHER K., KOEPLER O., KOHLHAMMER J., RUNNWERTH M., RUPPERT T., SCHRECK T., SENS I.: Visinfo: a digital library system for time series research data based on exploratory search - a user-centered design approach. *International Journal on Digital Libraries (IJoDL)* 16, 1 (2015), 37–59. doi:10.1007/s00799-014-0134-y. 7
- [BDV\*17] BERNARD J., DOBERMANN E., VÖGELE A., KRÜGER B., KOHLHAMMER J., FELLNER D.: Visual-interactive semi-supervised labeling of human motion capture data. In *SPIE Visualization and Data Analysis (VDA)* (2017). doi:10.2352/ISSN.2470-1173.2017.1.VDA-387. 2
- [Ber15] BERNARD J.: *Exploratory search in time-oriented primary data*. Dissertation, PhD, Technische Universität Darmstadt, Graphisch-Interaktive Systeme (GRIS), Darmstadt, Germany, 2015. URL: <http://tuprints.ulb.tu-darmstadt.de/5173/>. 2
- [BFG\*15] BÖGL M., FILZMOSER P., GSCHWANDTNER T., MIKSCH S., AIGNER W., RIND A., LAMMARSCH T.: Visually and statistically guided imputation of missing values in univariate seasonal time series. In *IEEE Visual Analytics Science and Technology (VAST)* (2015), pp. 189–190. doi:10.1109/VAST.2015.7347672. 2
- [BHJ\*14] BONNEAU G.-P., HEGE H.-C., JOHNSON C. R., OLIVEIRA M. M., POTTER K., RHEINGANS P., SCHULTZ T.: Overview and State-of-the-Art of Uncertainty Visualization. In *Scientific Visualization, Mathematics and Visualization*. Springer, 2014, pp. 3–27. URL: [https://link.springer.com/chapter/10.1007/978-1-4471-6497-5\\_1](https://link.springer.com/chapter/10.1007/978-1-4471-6497-5_1), doi:10.1007/978-1-4471-6497-5\_1. 3, 6
- [BKS12] BERNARD J., KÖNIG-LANGLO G., SIEGER R.: Time-oriented earth observation measurements from the Baseline Surface Radiation Network (BSRN) in the years 1992 to 2012, reference list of 6813 datasets, 2012. doi:10.1594/PANGAEA.787726. 7
- [BKS14] BERNARD J., KÖNIG-LANGLO G., SIEGER R.: 30 years of synoptic observations from Neumayer Station with links to datasets, 2014. doi:10.1594/PANGAEA.150017. 8
- [BPHE17] BOUKHELIFA N., PERRIN M.-E., HURON S., EAGAN J.: How Data Workers Cope with Uncertainty: A Task Characterisation Study. In *SIGRAD, Swedish Chapter of Eurographics* (2017), (New York, NY, USA, 2017), ACM, pp. 3645–3656. doi:10.1145/3025453.3025738. 3
- [BRG\*12] BERNARD J., RUPPERT T., GOROLL O., MAY T., KOHLHAMMER J.: Visual-interactive preprocessing of time series data. In *SIGRAD, Swedish Chapter of Eurographics* (2012), vol. 81, Linköping University Electronic Press, pp. 39–48. 3
- [BSH\*16] BACH B., SHI C., HEULOT N., MADHYASTHA T., GRABOWSKI T., DRAGICEVIC P.: Time curves: Folding time to visualize patterns of temporal evolution in data. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 22, 1 (2016), 559–568. doi:10.1109/TVCG.2015.2467851. 3
- [BSS\*18] BEHRISCH M., STREEB D., STOFFEL F., SEEBACHER D., MATEJEK B., WEBER S. H., MITTELSTAEDT S., PFISTER H., KEIM D.: Commercial visual analytics systems-advances in the big data analytics field. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* (2018), 1–1. doi:10.1109/TVCG.2018.2859973. 1, 4
- [BW08] BYRON L., WATTENBERG M.: Stacked graphs – geometry amp; aesthetics. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 14, 6 (2008), 1245–1252. doi:10.1109/TVCG.2008.166. 3
- [BWK\*13] BERNARD J., WILHELM N., KRÜGER B., MAY T., SCHRECK T., KOHLHAMMER J.: Motionexplorer: Exploratory search in human motion capture data based on hierarchical aggregation. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 19, 12 (2013), 2257–2266. doi:10.1109/TVCG.2013.178. 3
- [BWS\*12] BERNARD J., WILHELM N., SCHERER M., MAY T., SCHRECK T.: TimeSeriesPaths: Projection-Based Explorative Analysis of Multivariate Time Series Data. *Journal of WSCG* 20, 2 (2012), 97–106. URL: <http://wscg.zcu.cz/wscg2012/Index.htm>. 3
- [CCM09] CORREA C. D., CHAN Y., MA K.: A framework for uncertainty-aware visual analytics. In *IEEE Visual Analytics Science and Technology (VAST)* (2009), pp. 51–58. doi:10.1109/VAST.2009.5332611. 3
- [CGM\*17] CENEDA D., GSCHWANDTNER T., MAY T., MIKSCH S., SCHULZ H., STREIT M., TOMINSKI C.: Characterizing Guidance in Visual Analytics. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 23, 1 (2017), 111–120. doi:10.1109/TVCG.2016.2598468. 6
- [CLKS19] CORRELL M., LI M., KINDLMANN G., SCHEIDEGGER C.: Looks good to me: Visualizations as sanity checks. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 25, 1 (2019), 830–839. doi:10.1109/TVCG.2018.2864907. 3
- [DCE\*13] DEMŠAR J., CURK T., ERJAVEC A., ČRT GORUP, HOČEVAR T., MILUTINVIĆ M., MOŽINA M., POLAJNAR M., TOPLAK M., STARIĆ A., ŠTAJDOHAR M., UMEK L., ŽAGAR L., ŽBONTAR J., ŽITNIK M., ZUPAN B.: Orange: Data mining toolbox in python. *Journal of Machine Learning Research* 14 (2013), 2349–2353. URL: <http://jmlr.org/papers/v14/demsar13a.html>. 4
- [EA12] ESLING P., AGON C.: Time-series data mining. *ACM Computer Survey* 45, 1 (2012), 12:1–12:34. doi:10.1145/2379776.2379788. 2
- [EMJ10] ELMQVIST N., MCDONNELL B., JAVED W.: Graphical perception of multiple time series. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 16 (2010), 927–934. doi:10.1109/TVCG.2010.162. 3, 5, 6
- [FPS96] FAYYAD U. M., PIATETSKY-SHAPIRO G., SMYTH P.: From data mining to knowledge discovery in databases. *AI Magazine* 17, 3 (1996), 37–54. 2
- [Fu11] FU T.-C.: A review on time series data mining. *Engineering Applications of Artificial Intelligence* 24, 1 (2011), 164–181. doi:10.1016/j.engappai.2010.09.007. 2, 3
- [GAW\*11] GLEICHER M., ALBERS D., WALKER R., JUSUFI I.,

- HANSEN C. D., ROBERTS J. C.: Visual comparison for information visualization. *Information Visualization* 10, 4 (2011), 289–309. doi:10.1177/1473871611416549. 3
- [GBFM16] GSCHWANDTNER T., BÖGL M., FEDERICO P., MIKSCH S.: Visual Encodings of Temporal Uncertainty: A Comparative User Study. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 22, 1 (2016), 539–548. doi:10.1109/TVCG.2015.2467752. 3, 6
- [GE18] GSCHWANDTNER T., ERHART O.: Know your enemy: Identifying quality problems of time series data. In *IEEE Pacific Visualization Symposium (PacificVis)* (2018), pp. 205–214. doi:10.1109/PacificVis.2018.00034. 3
- [GGAM12] GSCHWANDTNER T., GÄRTNER J., AIGNER W., MIKSCH S.: A taxonomy of dirty time-oriented data. In *CD-ARES* (2012), vol. 7465 of *Lecture Notes in Computer Science*, Springer, pp. 58–72. 2
- [GTPB19] GOGOLOU A., TSANDILAS T., PALPANAS T., BEZERIANOS A.: Comparing similarity perception in time series visualizations. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 25, 1 (2019), 523–533. doi:10.1109/TVCG.2018.2865077. 3
- [HFH\*09] HALL M., FRANK E., HOLMES G., PFAHRINGER B., REUTEMANN P., WITTEN I. H.: The weka data mining software: An update. *ACM SIGKDD Explorations Newsletter* 11, 1 (2009), 10–18. doi:10.1145/1656274.1656278. 4
- [HHN00] HAVRE S., HETZLER B., NOWELL L.: Themeriver: visualizing theme changes over time. In *Symposium on Information Visualization (InfoVis)* (2000), IEEE, pp. 115–123. doi:10.1109/INFVIS.2000.885098. 3
- [HQC\*18] HULLMAN J., QIAO X., CORRELL M., KALE A., KAY M.: In Pursuit of Error: A Survey of Uncertainty Visualization Evaluation. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* (2018), 1–1. doi:10.1109/TVCG.2018.2864889. 3
- [HWX\*10] HU Y., WU S., XIA S., FU J., CHEN W.: Motion track: Visualizing variations of human motion data. In *IEEE Pacific Visualization Symposium (PacificVis)* (2010), IEEE, pp. 153–160. 3
- [jav] JAVAGL.DE: Flow. <https://github.com/javagl/Flow>. 4
- [KHP\*11] KANDEL S., HEER J., PLAISANT C., KENNEDY J., VAN HAM F., RICHE N. H., WEAVER C., LEE B., BRODBECK D., BUONO P.: Research directions in data wrangling: Visualizations and transformations for usable and credible data. *Information Visualization* 10, 4 (2011), 271–288. doi:10.1177/1473871611415994. 2, 3
- [KK03] KEOGH E., KASETTY S.: On the need for time series data mining benchmarks: A survey and empirical demonstration. *Springer Data Mining and Knowledge Discovery* 7, 4 (2003), 349–371. doi:10.1023/A:1024988512476. 2
- [KTWZ10] KRÜGER B., TAUTGES J., WEBER A., ZINKE A.: Fast local and global similarity searches in large motion capture databases. In *ACM SIGGRAPH/EG Symp. on Comp. Anim.* (2010), Eurographics, pp. 1–10. doi:10.2312/SCA/SCA10/001-010. 2
- [LK06] LARAMEE R. S., KOSARA R.: Challenges and unsolved problems. In *Human-Centered Visualization Environments* (2006), vol. 4417 of *Lecture Notes in Computer Science*, Springer, pp. 231–254. doi:10.1007/978-3-540-71949-6. 3
- [LKL\*04] LIN J., KEOGH E., LONARDI S., LANKFORD J. P., NYS-TROM D. M.: Visually mining and monitoring massive time series. In *ACM SIGKDD Knowledge Discovery and Data Mining (KDD)* (2004), ACM, pp. 460–469. doi:10.1145/1014052.1014104. 2
- [LS18] LIMA R., SAMPAIO R.: What is uncertainty quantification? *Journal of the Brazilian Society of Mechanical Sciences and Engineering* 40, 3 (2018), 155. doi:10.1007/s40430-018-1079-7. 3
- [LSDJ06] LEW M. S., SEBE N., DJERABA C., JAIN R.: Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 2, 1 (2006), 1–19. doi:10.1145/1126004.1126005. 2
- [LYK\*12] LUO D., YANG J., KRSTAJIC M., RIBARSKY W., KEIM D.: Eventriver: Visually exploring text collections with temporal references. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 18, 1 (2012), 93–105. doi:10.1109/TVCG.2010.225. 3
- [Mac15] MACEACHREN A. M.: Visual Analytics and Uncertainty: Its Not About the Data. In *EuroVis Workshop on Visual Analytics (EuroVA)* (2015), Eurographics. doi:10.2312/eurova.20151104. 3
- [MHK06] MOESLUND T. B., HILTON A., KRÜGER V.: A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding* 104, 2 (2006), 90–126. Special Issue on Modeling People: Vision-based understanding of a person's shape, appearance, movement and behaviour. doi:https://doi.org/10.1016/j.cviu.2006.08.002. 2
- [Mör06] MÖRCHEN F.: *Time series knowledge mining*. Citeseer, 2006. 2
- [MR06] MÜLLER M., RÖDER T.: Motion templates for automatic classification and retrieval of motion capture data. In *ACM SIGGRAPH/EG Symposium on Computer Animation (SCA)* (2006), Eurographics, pp. 137–146. 2
- [Mül07] MÜLLER M.: *Information Retrieval for Music and Motion*. Springer-Verlag New York, Inc., 2007. 2
- [NS12] NEWMAN G. E., SCHOLL B. J.: Bar graphs depicting averages are perceptually misinterpreted: The within-the-bar bias. *Psychonomic Bulletin & Review* 19, 4 (2012), 601–607. doi:10.3758/s13423-012-0247-5. 6
- [PWB\*09] POTTER K., WILSON A., BREMER P. T., WILLIAMS D., DOUTRIAUX C., PASCUCCI V., JOHNSON C. R.: Ensemble-Vis: A Framework for the Statistical Visualization of Ensemble Data. In *2009 IEEE International Conference on Data Mining Workshops* (2009), pp. 233–240. doi:10.1109/ICDMW.2009.55. 3
- [Rap] RAPIDMINER, INC: Rapidminer. <https://rapidminer.com/>. 4
- [RBS\*18] RAUTENHAUS M., BÖTTINGER M., SIEMEN S., HOFFMAN R., KIRBY R. M., MIRZARGAR M., RÖBER N., WESTERMANN R.: Visualization in meteorology—a survey of techniques and tools for data analysis tasks. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 24, 12 (2018), 3268–3296. doi:10.1109/TVCG.2017.2779501. 3
- [Rei08] REIJNER H.: The development of the horizon graph, 2008. 3
- [RLKLI12] RIMBU N., LOHMANN G., KÖNIG-LANGLO G., IONITA M.: 30 years of synoptic observations from Neumayer Station with links to datasets., 2012. 7, 8
- [SAAF18] SHURKHOVETSKYY G., ANDRIENKO N., ANDRIENKO G., FUCHS G.: Data abstraction for visualizing large time series. *Computer Graphics Forum (CGF)* 37, 1 (2018), 125–144. doi:10.1111/cgf.13237. 2
- [SHB\*14] SEDLMAIR M., HEINZL C., BRUCKNER S., PIRINGER H., MÖLLER T.: Visual parameter space analysis: A conceptual framework. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 20, 12 (2014), 2161–2170. doi:10.1109/TVCG.2014.2346321. 2, 3, 4, 5
- [SMF15] SAKURAI Y., MATSUBARA Y., FALOUTSOS C.: Mining and forecasting of big time-series data. In *ACM SIGMOD International Conference on Management of Data* (2015), ACM, pp. 919–922. doi:10.1145/2723372.2731081. 2
- [SS18] SONG H., SZAFIR D. A.: Where's my data? evaluating visualizations with missing data. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* (2018), 1–1. doi:10.1109/TVCG.2018.2864914. 2
- [SSK\*16] SACHA D., SENARATNE H., KWON B. C., ELLIS G., KEIM D. A.: The Role of Uncertainty, Awareness, and Trust in Visual Analytics. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 22, 1 (2016), 240–249. doi:10.1109/TVCG.2015.2467591. 2, 3
- [SSZ\*17] SACHA D., SEDLMAIR M., ZHANG L., LEE J. A., PELTONEN J., WEISKOPF D., NORTH S. C., KEIM D. A.: What you see is what you can change: Human-centered machine learning by interactive visualization. *Neurocomputing* (2017). ISSN = 0925-2312. doi:10.1016/j.neucom.2017.01.105. 5
- [SZS\*16] SACHA D., ZHANG L., SEDLMAIR M., LEE J. A., PELTONEN J., WEISKOPF D., NORTH S. C., KEIM D. A.: Visual Interaction with Dimensionality Reduction: A Structured Literature Analysis. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 23, 01 (2016), 241–250. doi:10.1109/TVCG.2016.2598495. 3, 9

- [vLFR17] VON LANDESBERGER T., FELLNER D. W., RUDDLE R. A.: Visualization system requirements for data processing pipeline design and optimization. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 23, 8 (2017), 2028–2041. doi:10.1109/TVCG.2016.2603178. 3,4
- [WBFL17] WUNDERLICH M., BALLWEG K., FUCHS G., LANDESBERGER T. v.: Visualization of Delay Uncertainty and its Impact on Train Trip Planning: A Design Study. *Computer Graphics Forum (CGF)* 36, 3 (2017), 317–328. doi:10.1111/cgf.13190. 3,6
- [WL05] WARREN LIAO T.: Clustering of time series data—a survey. *Pattern Recognition* 38, 11 (2005), 1857–1874. doi:10.1016/j.patcog.2005.01.025. 2
- [WVZ\*15] WILHELM N., VÖGELE A., ZSOLDOS R., LICKA T., KRÜGER B., BERNARD J.: Furyexplorer: Visual-interactive exploration of horse motion capture data. SPIE Press. doi:doi:10.1117/12.2080001. 3
- [WYM12] WU Y., YUAN G.-X., MA K.-L.: Visualizing flow of uncertainty through analytical processes. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 18, 12 (2012), 2526–2535. doi:10.1109/TVCG.2012.285. 2,3
- [ZCPB11] ZHAO J., CHEVALIER F., PIETRIGA E., BALAKRISHNAN R.: Exploratory analysis of time-series with chronolenses. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 17, 12 (2011), 2422–2431. doi:10.1109/TVCG.2011.195. 3
- [ZJGK10] ZIEGLER H., JENNY M., GRUSE T., KEIM D. A.: Visual market sector analysis for financial time series data. In *IEEE Visual Analytics Science and Technology (VAST)* (2010), IEEE, pp. 83–90. doi:10.1109/VAST.2010.5652530. 8