Visualizing Time Series Consistency for Feature Selection

Lena Cibulski Thorsten May
Fraunhofer IGD
Fraunhoferstr. 5
64283 Darmstadt, Germany
{first}.{last}@igd.fraunhofer.de

Bernhard Preim
OvGU Magdeburg
Universitätsplatz 2
39106 Magdeburg, Germany
bernhard.preim@ovgu.de

Jürgen Bernard Jörn Kohlhammer
TU Darmstadt
Fraunhoferstr. 5
64283 Darmstadt, Germany
{first}.{last}@gris.tu-darmstadt.de

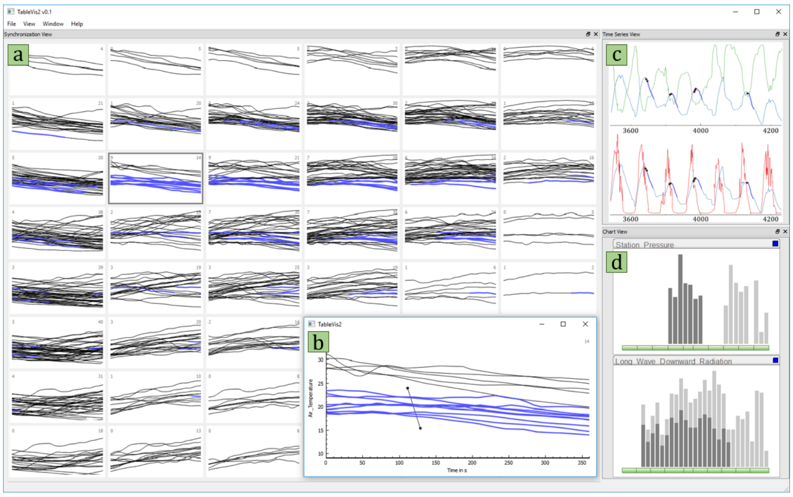


Figure 1: Visualizing how consistently a variable subset captures the characteristics of a dependent variable. Segments corresponding to groups of comparable values on the subset are investigated (a). An inconsistency is brushed (b, c). Histograms (d) show that one of the variables (top) might add to the variable subset's discriminating ability.

ABSTRACT

Feature selection is an effective technique to reduce dimensionality, for example when the condition of a system is to be understood from multivariate observations. The selection of variables often involves a priori assumptions about underlying phenomena. To avoid the associated uncertainty, we aim at a selection criterion that only considers the observations. For nominal data, consistency criteria meet this requirement: a variable subset is consistent, if no observations with equal values on the subset have different output values. Such a model-agnostic criterion is also desirable for forecasting. However, consistency has not yet been applied to multivariate time series. In this work, we propose a visual consistency-based technique for analyzing a time series subset's discriminating ability w.r.t. characteristics of an output variable. An overview visualization conveys the consistency of output progressions associated with comparable observations. Interaction concepts and detail visualizations provide a steering mechanism towards inconsistencies. We demonstrate the technique's applicability based on two real-world scenarios. The results indicate that the technique is open to any forecasting task that involves multivariate time series, because analysts could assess the combined discriminating ability without any knowledge about underlying phenomena.

Keywords
Visual Analytics, Feature Selection, Consistency, Multivariate Time Series, Model-Agnostic, Forecasting

1 INTRODUCTION

Predictive modeling is the process of analyzing historical observations to make a statement about the future. It is called *forecasting* when the time dimension is used as an explicit source of information. As an example, the missing values of a broken sensor can be expressed by a subset of remaining sensors in the network. With multiple sensors available, choosing a representative subset of input variables becomes a primary challenge. In machine learning, this is referred to as *feature selection*.

Feature selection techniques can be grouped according to the selection criterion used to rate a variable subset. Wrapper methods use a predictive model that is built from the subset. When the type of model can be determined, they usually yield well-performing variable subsets. However, solid indications that guide the decision for a model type might not be available. In such cases, filter methods, which explore data-based characteristics to evaluate a variable subset, might be more feasible.

Among the available filter criteria, most of which rely on combined bivariate or trivariate dependencies, *consistency* is the only one to combine being multivariate by design with two important benefits: *1)* it is model-agnostic, i.e. does not involve any assumption about the nature of data characteristics, and *2)* it removes both redundant and irrelevant variables, thus serving simplicity. Consistency has originally been defined for classification with nominal data. The idea is to identify a minimal variable subset that is consistent with a set of observations. A variable subset is inconsistent, if one or more pairs of observations exhibit equal values for the variables in the subset, but have different class labels. The more inconsistencies a variable subset induces, the worse its relevance for prediction.

However, consistency has not yet been considered in the context of forecasting. Most selection criteria for time series contain an implicit model specification, e.g. using principal components or pairwise cross correlation. This motivates us to investigate how far a *model-agnostic* evaluation of variable subsets for forecasting can get. We do not aim at a new modeling approach. Instead, we focus on feature selection as a preparatory step that guides the following modeling stages.

In this work, we propose a visual-interactive selection technique centered around a criterion that builds upon the consistency of temporal output progressions. To adapt the concept of consistency to time series, we establish counterparts for *1)* the similarity of observations and *2)* the similarity of corresponding outputs. Conceptually, we define an inconsistency as a pair of time points exhibiting similar value combinations with respect to the variable subset, but being followed by dissimilar temporal progressions of the output variable.

Our focus is on leveraging the capabilities of well-established visual analytics techniques to determine how consistently a variable subset captures multivariate time series data. For judging the similarity on the output side, we prefer visual perception over a quantified similarity measure to meet the data- and application-dependent understanding of similarity. The output progressions associated with comparable time points are plotted as aligned curves in a line chart. Perceived inconsistencies indicate portions of the output variable that the investigated variable subset cannot explain. We combine different compositions of line charts with interaction techniques to steer the analysis and to support a refinement of variable subsets (see Figure 1). The contributions of this paper include:

- A consistency-based, model-agnostic selection criterion for multivariate time series
- Overview visualizations conveying the consistency of output progressions across the subset domain
- Detail visualizations and interaction concepts that support the analysis and resolving of inconsistencies

2 RELATED WORK

Our visual-interactive approach intersects the areas of feature selection, regression and predictive modeling, and the analysis of time series. Our literature review shows a lack of feature selection techniques that are both model-agnostic and targeted at temporal data (Table 1). Instead, temporal dependencies are often embedded in similarity or correlation metrics.

2.1 Feature Selection

Molina et al. provide an overview of selection criteria and search strategies for feature selection [MBN02]. Blum and Langley distinguish feature selection methods based on the selection criterion [BL97]. In their terms, our work is most closely related to filters. Filter criteria are usually based on measures for distance, information, dependence, or consistency. Unlike the first three, where scalability is mostly achieved by combining bivariate dependencies, consistency can evaluate a subset of variables at a time, independent of its size. Dash and Liu find consistency to be fast and able to remove redundant variables [DL03]. Sips et al. were among the first to adopt consistency for 2D-projections of labeled numerical data [SNLH09]. However, the concept of consistency has not yet been adapted to time series, and we are aiming to close this gap.

Most visual approaches to feature selection are based on correlation measures. Guo details pairwise correlations in a matrix view to explore interesting variable subspaces for clustering [Guo03]. Krause et al.'s *SeekAView* enables a dual exploration of subspaces and item subsets [KDFB16]. *Infuse* by Krause et al. evaluates the performance of combinations of feature selection and classification techniques [KPB14], essentially integrating filter and wrapper approaches. The *Smartstripes* approach by May et al. turns an automated feature selection into a white-box approach, allowing the user to intervene in an automated heuristic [MBD$^+$11].

Feature selection techniques have also been applied to multivariate time series. Yoon et al. propose *CLeVer*, a ranking feature selection based on a principal component analysis [YYS05]. It effectively eliminates vari-

Method	Model-Agnostic	Time dependence	Partitioning	Expertise	Ranking
Our approach	✓	✓	✓	✓	
CLeVer [YYS05]		✓			✓
Incremental cross correlation [Bac16]		✓			✓
SmartStripes [MBD+11]	✓		✓	✓	✓
Partition-Based Framework [MP13]			✓	✓	✓
Interactive Feature Selection [Guo03]	✓			✓	
Infuse [KPB14]				✓	✓

Table 1: An overview of reviewed methods. Only our approach is model-agnostic *and* supports time dependence.

able redundancies, but does not consider time dependence. In contrast, Bacciu proposes a cross correlation approach that respects the temporal order [Bac16].

2.2 Regression Modeling and Forecasting

Predicting future values based on multivariate time series relates to aspects of both forecasting and regression. Regression models relationships among numerical variables. We contrast forecasting approaches that model relationships between different points in time. These two tasks actually complement each other and can often be performed with the same modeling techniques. Mühlbacher and Piringer present a comprehensive regression approach integrating variable ranking, localized and incremental adaption, and residual analysis [MP13]. In a recent approach, Matkovic et al. inject temporal dependencies into the regression model by means of scalar aggregates [MAJH17]. *TimeFork* by Badam et al. combine regression and forecasting represented by two neural networks for temporal and conditional predictions [BZS+16]. The link between the two models is continuously refined via user interaction and feedback. In principle, all of these approaches allow for choosing a model type, but the decision still has to be made prior to analysis and cannot be postponed.

2.3 Time Series Analysis

Our work is also related to visual analysis of time series ensembles. Konyha et al. present an ensemble visualization to analyze the influence of control parameters on the behavior of dependent time series [KMG+06]. In contrast, we observe how dynamic properties of input time series influence the dependent time series' behavior. Buono et al. propose an approach for the visual analysis of sample-based forecasts [BPS+07]. Like them, we filter variables to produce a comparable subset of samples, but we aim to systematically cover *all* potential filter settings. Schreck et al. use Self Organizing Maps to cluster time series, allowing for a scalable exploration of behaviors [SBvLK09]. However, the approach relies on the definition of a similarity metric, which requires assumptions about the underlying relations. The *TimeCurves* approach by Bach et al. reveals patterns of temporal evolution by distorting the temporal trajectory such that spatial proximity indicates similarity of events [BSH+16]. Instead of keeping the time series connected, we isolate chunks that belong to the same region of the variable subset domain.

Multivariate time series can be aligned to match with discrete events, providing a better frame of reference for comparisons. In *Lifeflow* [WGGP+11], discrete event series are aligned, while *CareCruiser* [GAK+11] aligns continuous time series to discrete events. Our approach also allows alignment to non-discrete events.

3 CONSISTENCY-BASED CRITERION

We first illustrate the concept of consistency-based feature selection on the basis of nominal data. We then explain the challenges of applying this concept to time series and describe our adapted consistency criterion.

3.1 Consistency for Nominal Data

A model approximates a function that maps some input variables $X = \{X_1, ... X_n\}$ to an output variable Y. The exact mapping is unknown, except for a collection of observations $(\vec{x}, y)$ with inputs $\vec{x} = (x_1, ..., x_n); x_i \in X_i$ and outputs $y \in Y$. Any attempt to approximate the mapping needs to capture the input-output assignments as specified in these observations. This requires that no input is associated with several different outputs. We thus search for a subset $X' \subset X$ of variables that fully distinguish among the outputs in the observations. We favor subsets defined over as few variables as possible, also referred to as *Min-Features bias* [AD91].

A variable subset X' allows for distinction of outputs, if no two observations with the same values on the subset have different outputs:

$$\forall \vec{x_1}', \vec{x_2}' \in X' : \vec{x_1}' = \vec{x_2}' \implies y_1 = y_2 \tag{1}$$

In contrast, outputs are not distinguishable by the variables in X', if two observations exhibit the same values on the subset, but are associated with different outputs:

$$\exists \vec{x_1}', \vec{x_2}' \in X' : \quad \underbrace{\vec{x_1}' = \vec{x_2}'}_{\text{input comparison}} \quad \wedge \quad \underbrace{y_1 \neq y_2}_{\text{output comparison}} \tag{2}$$

In this case, the output for input $\vec{x}_1{}' = \vec{x}_2{}'$ is ambiguous, thus introducing *inconsistency*. In the context of classification, an inconsistency is defined as "two instances [...] that are equal when considering only the [variables] in X' and that belong to different classes" [MBN02]. For deterministic models assigning exactly one output to each input, the ambiguity associated with inconsistencies always results in a modeling error. Consistency criteria therefore aim at inducing as few inconsistencies as possible using the smallest possible subset.

3.2 Adapting Consistency to Time Series

In this section, we describe our approach for transforming the concept of consistency into a selection criterion for multivariate time series data. Each observation $(\vec{x}, y)$ is now associated with a time stamp t. If the temporal order of these observations was ignored, they could be treated just like tabular data. However, this is not an option for an analysis of real-world scenarios.

Instead, we focus on dependencies between $\vec{x}$ and y that manifest across multiple time steps. In theory, the output value to be predicted at a certain time point might be affected by any input variable at any earlier time, although the effect of values might be larger for recent time points. Still, it is not feasible to search this space due to the huge number of variables and time points to be considered. Following the idea of consistency as described for nominal data (Section 3.1), such a dependency can be approached from a different perspective: inputs are considered as the starting point, whose effect on the output values at later times is investigated.

Just like before, we consider observations with the same values on the variable subset. Instead of classes as with nominal data, we now observe the dependent variable's temporal progressions $y : [t, t+l]$ over intervals $[t, t+l]$ on the output side. Thus, we re-define inconsistency as:

$$\exists t_a, t_b : \underbrace{\vec{x}'(t_a) \sim \vec{x}'(t_b)}_{\text{input comparison}} \wedge \underbrace{y : [t_a, t_a + l] \not\sim y : [t_b, t_b + l]}_{\text{output comparison}} \qquad (3)$$

Contrary to Equation (2), we consider input similarity rather than equality, because we cannot expect two observations to exhibit the exact same numbers for the numerical variables in the subset. Given Equation (3), a variable subset is said to be consistent if, by means of the inputs $\vec{x}'$ specified by the involved variables, dissimilar output progressions can be distinguished.

3.3 Analysis Tasks

To evaluate a variable subset using the adapted consistency criterion, Equation (3) is applied to multivariate time series. This is guided by the following questions:

- *Input comparison* – are numerical observations similar regarding the variable subset?
- *Output comparison* – are output progressions associated with similar inputs similar as well?
- *Inconsistency search* – where do outputs progress inconsistently across the variable subset domain?
- *Inconsistency analysis* – under which conditions do they occur? How can they be resolved?

Based on these leading questions, we propose the following tasks to be addressed by our visual technique to perform one iteration of the feature selection procedure:

T$_1$ Grouping of input vectors
T$_2$ Visualizing outputs for individual input groups
T$_3$ Layouting of visualizations for all input groups
T$_4$ In-depth analysis of inconsistencies

4 VISUAL FEATURE SELECTION

In the following, we present the role of interactive visualization to address the tasks of consistency-based feature selection as explained in Section 3. They apply to the subset evaluation and candidate generation in each iteration of the procedure. Note that the increasing dimensionality of the variable subset affects the part of an observation that we consider as input.

Both input and output comparisons require an understanding of similarity. For the input space, we present a grouping strategy that subdivides the value domain of a variable subset into disjoint regions of similar inputs (Section 4.1). The actual challenge lies in representing the task-dependent understanding of output similarity. The output progressions for each region are aligned for visual comparison in a dedicated chart (Section 4.2). The inconsistency across regions is conveyed by small multiples arranged in 2D visual space (Section 4.3). Finally, we provide interaction concepts for in-depth analysis and elimination of inconsistencies (Section 4.4).

4.1 Grouping

As described in Section 3.2, a variable subset is evaluated in terms of its consistency. It involves the accumulation of inconsistencies for all possible expressions of inputs as occurring in the data. Before we can assess inconsistencies, we thus need to perform a *grouping* of inputs to unique expressions (**T**$_1$).

A straight-forward solution to grouping is an equidistant binning strategy as it is known from histogram generation. For two or more variables, we use a regular grid, which corresponds to the Cartesian product of the individual binnings. The binning resolution can be adjusted at any point of the analysis. A finer resolution represents the input domain more accurately, but might lead to more regions being empty or containing a statistically insignificant number of time points. Regardless of the number of variables, the result of grouping are sets of time points where inputs are highly similar.

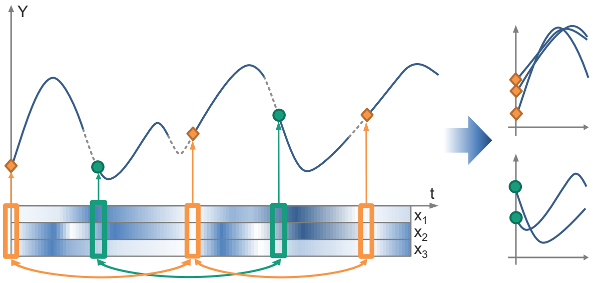


Figure 2: Visual comparison of output progressions. Synchronization points mark comparable inputs (left). The output progressions are cropped and shifted along the time axis to establish a common baseline (right).

We choose binning as a simple, yet effective solution, because it does not involve assumptions. Additional benefits include: *1)* axis-orthogonal partition boundaries are more easily translated into a visual representation than free-form regions, *2)* axis-orthogonal boundaries allow for a description of individual regions by means of value ranges, and *3)* it is free of unnecessary functionality such as maximizing inter-group distances.

4.2 Visualization of Individual Groups

The binning yields a discretization of the variable subset domain according to the similarity of observations. The purpose of visualization is to expose patterns of those output progressions that correspond to each group of similar inputs ($\mathbf{T}_2$). To make the progressions comparable, we propose the following alignment technique.

Synchronization The synchronization takes inputs that share common characteristics with respect to the variable subset and aligns the corresponding output progressions. It is performed in three steps (Figure 2):

1) Find all time points at which the given inputs arise
2) Split the output time series at these points and crop the sections to an interval length l
3) Shift the cropped sections on top of each other

Comparable inputs arise at different time points t_i, which we call *synchronization points* (Figure 2, left). As noted in Equation (3), the outputs to be compared are the progressions $y : [t_i, t_i + l]$ that evolve from the synchronization points onward. Under the premise that all t_i represent a common state, consistent progressions should be similar in relation to the synchronization points. We establish a common baseline by shifting the t_i as well as the progressions bound to them along the time axis, such that the synchronization points are mapped to $t = 0$. This is what we call *synchronization*. The result is depicted in Figure 2, right.

Synchronization Chart The *Synchronization Chart* depicts the output progressions obtained from the synchronization as an ensemble of line segments ($\mathbf{T}_2$). In contrast to traditional line charts, the time axis does not represent absolute times but relative to the synchronization points. The chart allows analysts to gain insight into inconsistencies among the output progressions. What is considered similar may vary depending on the nature of the data, the application, and the analysis focus. Providing an adequate visual representation allows analysts to take advantage of their visual understanding of similarity.

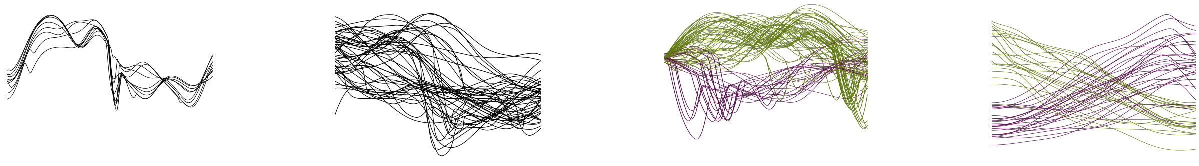
Figure 3 shows a collection of patterns indicating different forms of the relation between a variable subset and the output variable. The most meaningful patterns to be observed are highly similar output progressions (Figure 3a) or arbitrary behavior (Figure 3b). These two cover the entire spectrum from consistency to inconsistency. Another example are two diverging sets of progressions (Figure 3c). The common starting point indicates a relation between variable subset and output at least at the synchronization point. However, the subset seems to be missing a variable that discriminates between positive (green) and negative (purple) slope of the progressions. Another example are contrary behaviors (Figure 3d). The symmetry of the two groups and the uniformity of progressions within a group indicate that the variable subset already captures the output behavior quite well. The remaining separation of the groups might be done based on the output values at the synchronization point.

4.3 Layouting

For each non-empty group of inputs resulting from binning, the synchronized output progressions are visualized in a chart. What is missing for a broad assessment of consistency across the variable subset domain is a layouting strategy that arranges all charts in an overview visualization ($\mathbf{T}_3$). This strategy needs to cope with a common problem in the visualization of multivariate data: an arrangement in 2D visual space cannot preserve the positions of the groups in nD space.

To perceive how consistently a variable subset captures the output characteristics, it is more important to observe the spread of progressions within each group of inputs than between groups. This allows for a simplification of the visualization at two levels: *1)* the presentation of charts and *2)* their arrangement. For the presentation, we omit axes and labels of the charts to reduce visual clutter and to address the limited screen space available for each chart. Regarding the arrangement of the resulting small multiples, small errors in the representation of the topology in n-dimensional subset space are acceptable, as the focus lies on within-group comparisons of progressions. Thus, we trade a non-exact topology preservation for an efficient arrangement of charts, meaning simple and not costly to compute.

Finally, the layouting strategy is determined by the number of variables in the subset. For one or two variables, the input space can be directly mapped to



(a) Similar behaviors (b) Arbitrary behaviors (c) Diverging behaviors (d) Contrary behaviors

Figure 3: Patterns of output progressions convey different notions of consistency: similar progressions indicate consistency (a), arbitrary output behaviors indicate inconsistency (b), a shared origin indicates consistency at least at the synchronization point (c), distinguishable behaviors indicate the need for a discriminating variable (c,d).

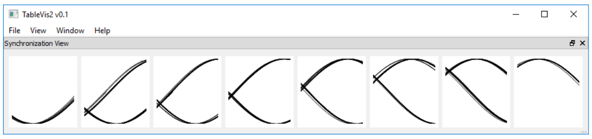


Figure 4: Synchronization Charts depict the output progressions across the value domain of a single variable. Despite the divergence, similarity indicates consistency.

the visual space. For one variable, charts are concatenated to a *Synchronization Row* (Figure 4), representing the intervals that result from binning. For two variables, the charts are arranged in a *Synchronization Grid* (Figure 5), which corresponds to the regular grid used for binning. Higher-dimensional subsets require additional considerations concerning the reduction to the two-dimensional visual space. The most pragmatic layouting strategy is to use dimension reduction. Even if the topology of inputs cannot be completely preserved, a layout based on dimension reduction might help to steer the analysis in a meaningful way.

4.4 In-Depth Analysis of Inconsistencies

The overview visualization enables analysts to perceive regions, where output progressions are dissimilar. However, the mere existence of inconsistencies does not always indicate an incomplete variable subset. An inconsistency may occur for two reasons: either *1)* the variable subset is missing some relevant variables or *2)* the underlying inputs, too, vary over time. Inconsistencies caused by the latter are considered false positives. The next step towards a refinement of the variable subset is to examine the circumstances under which the inconsistency occurs to eliminate false positives ($\mathbf{T}_4$).

Focus Visualization and Brushing To avoid similarity metrics wherever possible, our technique heavily relies on an interactive definition of what is perceived as (dis-)similar. Having identified an inconsistency in the overview visualization, analysts can bring the corresponding small multiple to focus. It then turns into an enlarged Synchronization Chart with axes and labels for orientation (Figure 5, center). This *Focus Plot* allows for interaction like zooming, panning, and brushing to investigate output progressions in more detail, while keeping the variable subset domain as a context.

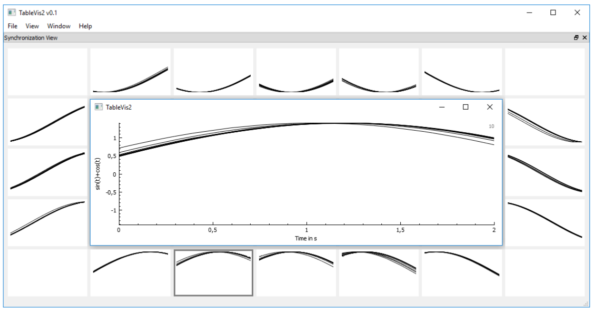


Figure 5: 2D variable subsets are visualized by small multiples in a regular grid. An enlarged chart allows for in-depth analysis of progressions in a region of interest.

From the detailed exploration, the analyst might have identified a number of progressions that make up an inconsistency. To investigate related data characteristics, the corresponding curves are selected. We provide a line brush (Figure 6a) as well as a rectangular brush. During analysis, users can flexibly adjust the mode. Brushing output progressions defines sequences of absolute time points, on which any further analysis is focused. In the following, we describe interaction concepts (Figure 6) that allow analysts to explore different aspects of any perceivable pattern of progressions.

Temporal Context In a Synchronization chart, all output progressions are depicted relative to the synchronization points. An overall temporal context is not given. For analysts, it requires significant cognitive effort to build up a mental image of how the depicted progressions are actually part of the same output time series. Viewing synchronization points and output progressions in a temporal context raises awareness for inappropriate default choices or inconclusive patterns in the output progressions that may otherwise be misinterpreted. We support analysts by linking brushed output progressions to a line chart with respect to *1)* the synchronization points and *2)* the actual progressions (Figure 6b). For *1)*, markers at the respective absolute time coordinates in the line chart indicate the synchronization points. For *2)*, the corresponding sequences of absolute time points are highlighted in the line chart.

Analysis of Input Variations Some output inconsistencies can be explained by variations of the underlying

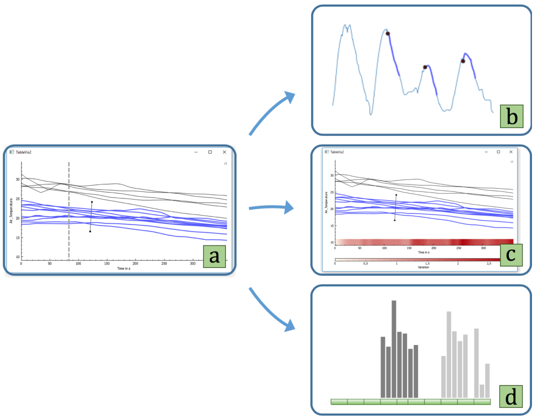


Figure 6: (a) Interaction in the Focus Plot: (b) brushed sections show in the global time range, (c) a one-row heatmap depicts the deviation of inputs, (d) brushed inputs are shown in histograms for candidate generation.

inputs. To distinguish these false positives from true inconsistencies, we investigate how the variance of inputs develops along the relative time points in a synchronization interval. Each relative time point is associated with a number of absolute time points. We aim at validating that the input similarity remains unchanged.

This can be performed based on aggregation using the *Variation Indicator*. It is a one-dimensional heat map along the time axis of a Synchronization Chart (Figure 6c). Color depicts the standard deviation of inputs at each relative time coordinate. Little and, if any, slowly increasing deviation indicates that input variation cannot be accounted for variations in the output. Consequently, any conclusion that is drawn from inconsistencies can be focused on variable subset incompleteness. The indicator can be consulted at any stage of the analysis to exclude output inconsistencies that are not necessarily caused by variable subset incompleteness.

Resolving Inconsistencies An inconsistency can be resolved by adding a variable that contributes to a separation between brushed and non-brushed progressions. The best candidates are variables whose values differ exactly where the output progressions differ. For diverging progressions, this could be a variable whose low and high values co-occur with the brushed and non-brushed progressions respectively. Investigating how progressions relate to the values of a variable provides valuable hints on its suitability as additional predictor.

To determine the relation of a candidate variable X_c to a brush, we compare two value distributions:

- *Focus distribution*: those values of X_c that are associated with the brushed observations
- *Context distribution*: those values of X_c that are associated with all observations depicted in the cell

If the shape of the focus distribution follows that of the context distribution, candidate variable and brushed progressions are considered independent (Table 2, left).

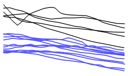
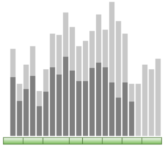
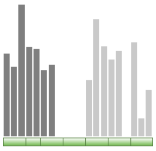
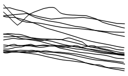
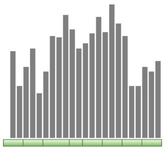
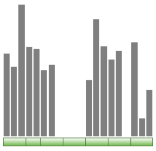
		Separation	
		No	Yes
Focus			
Context			

Table 2: Visually assessing the relation of a variable to a brush. Similar distributions indicate that there is none (left). Large differences suggest that the variable separates brushed from non-brushed progressions (right).

Relating inconsistencies to the value characteristics of a candidate variable is supported by depicting the value distributions in a histogram (Figure 6d). If the focus distribution highly differs from the context distribution (Table 2, right), this indicates that the candidate variable is related to the brushed progressions. It means that the variable carries significant potential to separate the brushed progressions from the rest. Including it into the current variable subset might lead to a better discrimination of the output, thus increasing the consistency.

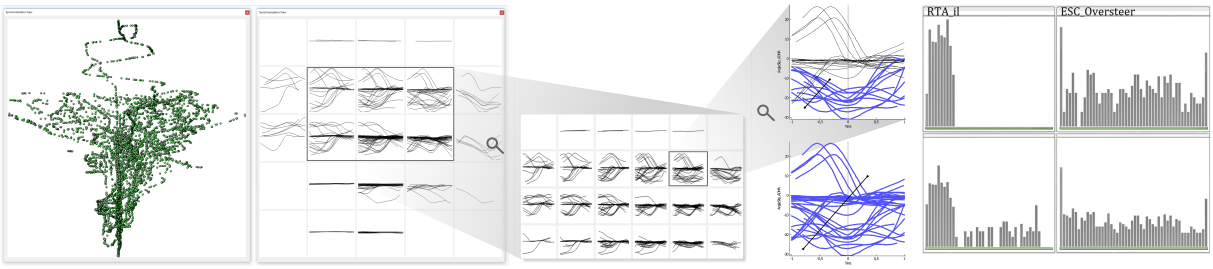
5 PROOF OF CONCEPT

We outline the benefits and limitations of our technique by performing one feature selection iteration using two different multivariate time series data sets. This proof of concept illustrates what can be learned from visualizations and interaction at each stage of the work flow. It aims at demonstrating the potential of visually perceived consistency as a model-agnostic criterion.

5.1 Real-World Meteorology Data

Meteorologists study weather events and climate trends to explain how the atmosphere affects the earth and to predict future weather developments. To investigate the effects of radiation and weather measurements on the temperature, we use data of the Baseline Surface Radiation Network [Beh11]. The data set shown in Figure 1 contains temperature, humidity, air pressure, and radiation measurements from August 2003, when Europe experienced an exceptionally hot summer.

Temperature is defined as the output variable. An initial variable subset is determined from visual exploration of line charts. We observe a negative correlation between *temperature* and *humidity* (Figure 1c, top). Furthermore, we notice that *shortwave-downward radiation* is also related to the *temperature* (Figure 1c, bottom). According to the domain experts, this relation can be explained with cloud-cover vs. clear-sky conditions.



(a) (b) (c) (d) (e)

Figure 7: A feature selection iteration given a 2D variable subset. An overview of the input domain (a) underlying the Synchronization Grid (b) supports the perception of consistency. Increasing the resolution for an inconsistent region hints at the most critical cell (c). The search for additional predictors is driven by distinguishable sets of output progressions (d). Histograms suggest that *RTA_il* might increase consistency when added to the subset (e).

Figure 1a shows the *temperature* progressions evolving from different combinations of *shortwave-downward radiation* and *humidity* within an interval of about six hours. The progressions do not behave arbitrarily, exhibiting an overall consistency across the domain. However, there are also regions of inconsistency exhibiting two distinguishable behaviors (Figure 1b).

Adding a third variable that separates these two types might significantly increase the explanatory power of the current variable subset. Based on the suggestions of the domain experts, we consider *longwave-downward radiation* and *station pressure* as candidates. For in-depth analysis, we brush one of the two separable sets of *temperature* progressions (Figure 1b). For the temporal context, the brushed sections and the synchronization points are highlighted in the line charts (Figure 1c).

Given the brushed set of *temperature* progressions, the histograms for *longwave-downward radiation* and *station pressure* show the corresponding focus distributions in dark-gray, while context distributions are depicted in light-gray (Figure 1d). The differences between focus and context distributions regarding shape and domain coverage in either histogram indicate that both candidate variables are related to the brush. *Station pressure* even achieves a perfect discrimination of the *temperature* behaviors (Figure 1d, top). It is therefore a promising variable to refine the variable subset. According to the experts, *station pressure* reflects the weather conditions connected to the atmosphere, thus relating to *temperature* beyond the daily periodicity of measurements. This confirms our findings.

Note, that these conclusions are drawn from analysis of a small part of the variable subset domain. An outlook of the consistency of the newly generated variable subset could verify the choice before proceeding with the next iteration of the feature selection.

5.2 Real-World Car Sensor Data

We also applied our technique to sensor data from the automotive domain. The data was acquired in the context of vehicle dynamics simulation [Unt13]. Due to a malfunction, the values of the *slip angle* sensor needed to be expressed by a subset of the remaining sensors.

Based on their domain knowledge, the experts suggested two variables to start the analysis with: *yaw rate* and *velocity*. To gain an overview of their value domain, we first take a look at clusters, empty regions, and outliers in a scatter plot (Figure 7a). A first notion of how the output variable behaves across this domain is conveyed in a Synchronization Grid with an initial binning resolution of 5×5 (Figure 7b). Despite a low level of detail, analysts can identify whether output progressions follow certain patterns or behave arbitrarily.

The analysis is driven by regions where output progressions are dissimilar, indicating that – at least for this particular part of the domain – the initial variables do not yet consistently capture all output characteristics. To fix this, we need to investigate the circumstances under which the inconsistency occurs. The drill-down to the most critical cell for detailed analysis requires an increased the grid resolution (Figure 7c).

Adding the variable that most effectively separates the depicted output behaviors might increase the discriminating ability of *{yaw rate,velocity}*. Based on domain knowledge or prior findings, candidate variables have been identified. To verify their suitability as additional predictors, we examine their relation to the brushed output progressions (Figure 7d, top) and the entirety of curves (Figure 7d, bottom) as described in Section 4.4.

For the first variable, the value distribution associated with the brush clearly deviates from the overall distribution (Figure 7e, left), not only in shape but also in domain coverage. Its separating ability makes it a promising variable to be added to the current variable subset. In contrast, the value distribution of the second variable is only marginally influenced by brushing the subset of output progressions (Figure 7e, right). This suggests that it does not provide separating potential and is thus unsuitable as a predictor. Still, such conclusions need to be verified in collaboration with domain experts.

6 DISCUSSION

The most important benefit of our visual feature selection technique is its generality. After selection, the typical decisions to be made for modeling, e.g. model type and parameters, are still available for disposition. In the common case that a model cannot be specified in advance, it provides a meaningful selection of variables rather than searching among different combinations of models and variable subsets in an exhaustive manner.

Unlike most forecasting methods, the proposed criterion does not consider historical patterns on the input side, which might significantly enhance the prediction of output progressions. We plan to address such an extension to the model-agnostic time series criterion in our future research. This might also involve considerations concerning an application to streaming data.

Due to its generality, our technique can be applied to any feature selection task involving multivariate time series. To convey the essential idea of our proposed criterion, we informally evaluated it by means of use cases from the meteorology and automotive domain. The proof of concept showed that, for one iteration, the technique effectively supported the subset evaluation and candidate generation. However, the limitations of our technique, in particular compared to alternative methods, need to be examined by a further evaluation.

While the proof of concept in this paper refers to the meteorology and automotive domain, preliminary tests with artificial data sets also resulted in useful findings. Even though the variable subset had been initialized with distracting variables, the true predictors could be identified in the further course of the feature selection. However, the technique did not produce the minimum subset. The uncertainty about the optimality of the solution is a limitation of virtually all heuristic approaches and our work is no exception.

To fully exploit the multivariate nature of our proposed time series criterion, the input grouping and chart layouting, too, should be multivariate. The Cartesian products of individual variable binnings used for grouping might result in empty or extremely dense regions that impair a statistical analysis of inconsistencies. Additionally, similar inputs might end up on two sides of an interval boundary. Both could be avoided by using adaptive grouping approaches that flexibly determine interval boundaries according to an objective function. The chart layouting is a more challenging task. In principle, any deterministic layout allows for assessing the consistency of a variable subset, even if charts are positioned randomly. However, a topology-preserving layout allows to associate perceived inconsistencies with the depicted portion of the variable subset domain. Approaches like dimension reduction could be combined with force-directed layouts to prevent overlaps. Although our pragmatic strategies could be considered as a limitation for now, both components, the grouping and the layouting, are open to any (not yet realized) scalable technique that suits the data and analysis best.

7 CONCLUSION & FUTURE WORK

In this paper, we presented a visual-interactive consistency-based feature selection technique for multivariate time series data. A key benefit of our technique is the purely model-agnostic evaluation of the combined quality of a variable subset. The criterion is solely based on the input-output characteristics reflected in time-dependent observations. It describes how consistently the subset captures these characteristics. The comparison of temporal progressions necessary for that purpose is solved visually to avoid uncertain assumptions about the nature of the data. Small multiples depict progressions that evolve from comparable inputs across the variable subset domain. An input grouping and layouting strategy present the charts in a way that allows for a detailed analysis of inconsistencies, working toward a refinement of the variable subset under consideration of previous insights. This is supported by interaction concepts that allow analysts to focus on findings, contribute their domain knowledge, and exclude false positive consistencies. Our technique is applicable as a preparatory step to any modeling task involving multivariate time series, in particular when decisions about the model specification shall be postponed.

We identify several directions for future work. The pragmatic approaches to grouping of inputs and layouting of charts could be extended by more scalable techniques to achieve a better representation of higher-dimensional variable subsets. A search strategy dedicated to work with our proposed criterion might provide a valuable contribution to reaching the goal of a minimal variable subset inducing the fewest possible inconsistencies. The strategy might also involve user guidance in the form of a ranking of recommended additional predictors. Finally, we plan to examine the potential of our technique for the validation of existing forecasting models as well as for data quality control. Inconsistencies regarding inputs and model outputs might indicate unsuitable parameter choices or an incomplete variable subset, and thus indicate a questionable model. Inconsistencies might also hint at low data quality, e.g. implausible spikes. In the case of missing data, our technique might help to impute values that are consistent with the observed data.

8 ACKNOWLEDGMENTS

This work was funded by the EC H2020 programme under grant agreement n° 768892.

9 REFERENCES

[AD91] Hussein Almuallim and Thomas G Dietterich. Learning with many irrelevant features. In *Association for the Advancement of Artificial Intelligence (AAAI)*, vol. 91, pp. 547–552, 1991.

[Bac16] Davide Bacciu. Unsupervised feature selection for sensor time-series in pervasive computing applications. *Neural Computing and Applications*, 27(5):1077–1091, 2016.

[Beh11] Klaus Behrens. Basic measurements of radiation at station lindenberg (2003-08), 2011.

[BL97] Avrim L Blum and Pat Langley. Selection of relevant features and examples in machine learning. *Artificial intelligence*, 97(1):245–271, 1997.

[BPS+07] Paolo Buono, Catherine Plaisant, Adalberto Simeone, Aleks Aris, Galit Shmueli, and Wolfgang Jank. Similarity-based forecasting with simultaneous previews: A river plot interface for time series forecasting. In *11th Int. Conf. on Information Visualization*, pp. 191–196, 2007.

[BSH+16] Benjamin Bach, Conglei Shi, Nicolas Heulot, Tara Madhyastha, Tom Grabowski, and Pierre Dragicevic. Time curves: Folding time to visualize patterns of temporal evolution in data. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):559–568, 2016.

[BZS+16] Sriram Karthik Badam, Jieqiong Zhao, Shivalik Sen, Niklas Elmqvist, and David Ebert. Timefork: Interactive prediction of time series. In *Proc. of the CHI Conference on Human Factors in Computing Systems*, pp. 5409–5420, 2016.

[DL03] Manoranjan Dash and Huan Liu. Consistency-based search in feature selection. *Artificial Intelligence*, 151(1-2):155–176, 2003.

[GAK+11] Theresia Gschwandtner, Wolfgang Aigner, Katharina Kaiser, Silvia Miksch, and Andreas Seyfang. Carecruiser: Exploring and visualizing plans, events, and effects interactively. In *IEEE Pacific Visualization Symposium (PacificVis)*, pp. 43–50, 2011.

[Guo03] Diansheng Guo. Coordinating computational and visual approaches for interactive feature selection and multivariate clustering. *Information Visualization*, 2(4):232–246, 2003.

[KDFB16] Josua Krause, Aritra Dasgupta, Jean-Daniel Fekete, and Enrico Bertini. Seekaview: An intelligent dimensionality reduction strategy for navigating high-dimensional data spaces. In *IEEE Symposium on Large Data Analysis and Visualization (LDAV)*, 2016.

[KMG+06] Zoltan Konyha, Kresimir Matkovic, Denis Gracanin, Mario Jelovic, and Helwig Hauser. Interactive visual analysis of families of function graphs. *IEEE Transactions on Visualization and Computer Graphics*, 12(6):1373–1385, 2006.

[KPB14] Josua Krause, Adam Perer, and Enrico Bertini. INFUSE: Interactive feature selection for predictive modeling of high dimensional data. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1614–1623, 2014.

[MAJH17] Krešimir Matković, Hrvoje Abraham, Mario Jelović, and Helwig Hauser. Quantitative externalization of visual data analysis results using local regression models. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, pp. 199–218, 2017.

[MBD+11] Thorsten May, Andreas Bannach, James Davey, Tobias Ruppert, and Jörn Kohlhammer. Guiding feature subset selection with an interactive visualization. In *IEEE Conference on Visual Analytics Science and Technology (VAST)*, 2011.

[MBN02] Luis Carlos Molina, Lluís Belanche, and Àngela Nebot. Feature selection algorithms: A survey and experimental evaluation. In *Proc. of the IEEE Int. Conf. on Data Mining*, pp. 306–313, 2002.

[MP13] Thomas Mühlbacher and Harald Piringer. A partition-based framework for building and validating regression models. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):1962–1971, 2013.

[SBvLK09] Tobias Schreck, Jürgen Bernard, Tatiana von Landesberger, and Jörn Kohlhammer. Visual cluster analysis of trajectory data with interactive kohonen maps. *Information Visualization*, 8(1):14–29, 2009.

[SNLH09] Mike Sips, Boris Neubert, John P. Lewis, and Pat Hanrahan. Selecting good views of high-dimensional data using class consistency. In *Computer Graphics Forum*, vol. 28, pp. 831–838, 2009.

[Unt13] Michael Unterreiner. *Modellbildung und Simulation von Fahrzeugmodellen unterschiedlicher Komplexität*. PhD thesis, Universität Duisburg-Essen, 2013.

[WGGP+11] Krist Wongsuphasawat, John Alexis Guerra Gómez, Catherine Plaisant, Taowei David Wang, Meirav Taieb-Maimon, and Ben Shneiderman. Lifeflow: Visualizing an overview of event sequences. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1747–1756, 2011.

[YYS05] Hyunjin Yoon, Kiyoung Yang, and Cyrus Shahabi. Feature subset selection and feature ranking for multivariate time series. *IEEE Transactions on Knowledge and Data Engineering*, 17(9):1186–1198, 2005.